DANGEROUS PREDICTIONS: EVALUATION METHODS FOR AND CONSEQUENCES OF PREDICTING DANGEROUS BEHAVIOR.

BY

EHSAN BOKHARI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Professor Lawrence Hubert, Chair
Professor Carolyn J. Anderson
Professor Steven A. Culpepper
Professor Fritz Drasgow
Professor James Rounds

Abstract

This thesis focuses on prediction in the social sciences. We begin by discussing the "clinical efficiency" of prediction methods as defined by Meehl and Rosen (1955), and present three equivalent conditions for assessing clinical efficiency: (1) a probabilistic inequality from Meehl and Rosen; (2) an equivalent inequality given by Dawes (1962); and (3) a more flexible and easily computed inequality that we refer to as the Bokhari-Hubert condition. The misuse of the area under the receiver operating characteristic (ROC) curve (AUC) is discussed, particularly when base rates are low. The biases associated with the AUC are examined with a recommendation that the positive and negative predictive values deserve more emphasis than typically provided in the literature. A thorough review is given for cross-validation, an important but often ignored evaluation strategy in developing predictive models. The bias-variance trade-off in prediction is explained, and several shrinkage estimators are examined. To facilitate our discussion, illustrative examples using predictive methods in criminology are provided and several are extensively examined. A detailed history of predicting dangerous and violent behavior is also given. As a final conclusion, great caution should be exercised when predicting outcomes with serious social justice consequences.

To Katherine,

the only predictable thing about you is your constant support.

Acknowledgments

First and foremost I would like to acknowledge my committee chairman and academic advisor, Lawrence Hubert. Prof. Hubert provided me with thoughtful guidance and expertise throughout the work of this thesis. He made an academic out of me, and for this I owe him my utmost gratitude.

My committee members Profs. Carolyn Anderson, Steve Culpepper, Fritz Drasgow, and Jim Rounds deserve great appreciation for all their insightful comments and suggestions. I would like to especially recognize Prof. Drasgow for providing his expertise during the preliminary work for the qualification exams and the two papers that were eventually included in this thesis as Chapters 3 and 4.

A special thank you goes to Sungjin Hong who provided invaluable guidance during his time as my advisor. It is unfortunate that he did not remain at the University of Illinois to see me through my graduate school tenure, but I know that he has moved on to a fulfilling career.

In addition to the academic support I received as a graduate student, I must also acknowledge the professors in the Departments of Mathematics and Psychology at the University of Arizona; in particular, Lee Sechrest and David Sbarra for their mentorship. Their support given to me in my final year at the U of A helped shape me into the future graduate student I became. I have been blessed with many influential and supportive educators throughout my educational career. All the teachers who went above and beyond deserve my appreciation.

Clearly I have been blessed with wonderful academic support, but just as important

is the emotional support that I received from my friends and family. My friends from the University of Illinois and from my childhood, particularly Noah and Cyrus, deserve much appreciation for their encouragement throughout my academic career. My family—especially my parents, my sister Adeela, and two brothers Jameel and Imraan—always supportive with praise, deserve plenty of thanks and love.

I owe the deepest gratitude to my beautiful fiancé, Katherine. Her support began from the moment we met and remained strong throughout. It would be cliché to say that I could not have succeeded in graduate school without her, but I do know it certainly would have been a lot more difficult if I did not have the love and support that she has given (as well as some thorough editing of Chapter 2).

Finally, I would like to mention those who have suffered the injustices of the criminal system and the victims of the heinous crimes committed by violent individuals. It is my hope that the topics presented in this thesis will lead to better violence prediction methods so that innocent people are not locked away and the lives of potential victims are saved.

Contents

Chapter 1

Introduction

"Prediction is very difficult, especially about the future."

— Niels Bohr

On September 16, 2013, Aaron Alexis killed thirteen people and injured eight others at the Washington Navy Yard in Southeast Washington, D.C. It was later reported that Alexis was "delusional" and told police in Rhode Island that he was hearing voices (Botelho & Sterling, 2013). On December 14, 2012, Adam Lanza shot and killed his mother before killing twenty children and six adults at Sandy Hook Elementary School in Newton, Connecticut. According to one report (Llanos, 2012) Lanza's older brother claimed Adam had a "history of mental problems" (para. 3). On July 20, 2012 in Aurora, Colorado, James Holmes opened fire in a crowded movie theater premiering *The Dark Knight Rises*, killing twelve and injuring fifty-eight more. Following the shooting it was noted that Holmes had previously seen "at least three mental health professionals" (Sallinger, 2012, para. 1). On January 8, 2011, Jared Loughner killed six and injured thirteen outside of a supermarket in Tucson, Arizona, including then-Congresswoman Gabrielle Giffords. Several reports followed suggesting that Loughner was severely mentally ill (e.g., Cloud, 2011).

On November 9, 2012a, the magazine *Mother Jones* published an article on its website entitled "Mass Shootings: Maybe What We Need Is a Better Mental-Health Policy" (Follman). In the article, Follman noted that many mass shooting suspects have had mental health issues: "Mass shootings generate sensational media coverage, yet most media have failed to connect the dots with regard to mental health" (para. 5). "Connect the dots" is exactly what Follman, Aronsen, Pan, and Caldwell (2012) attempted to do; on the *Mother*

Jones website the authors made data available regarding sixty-seven different mass shootings in the United States since 1982, including the four mentioned in the opening paragraph (it is continually updated following new cases). Among other details, they provide the number of victims (fatalities and injuries), the suspect's race and sex, the type of weapon(s) used and if they were legally obtained or not, and whether the suspect had any prior signs of mental illness. They concluded that forty-two of the killers—about 63%—may have displayed signs of mental illness prior to their crimes.

Following the so-called Navy-Yard Shooting, Denise Grady from the *New York Times* wrote an article entitled "Signs May Be Evident in Hindsight, but Predicting Violent Behavior Is Tough" (2013). She questions whether an incident such as the Navy-Yard Shooting could be predicted. One of her interviewees was Dr. Jeffrey W. Swanson from Duke University's Department of Psychiatry. When discussing the possibility of predicting such a mass shooting, he commented, "I can tell you the common characteristics of people who engage in mass shootings: It's a picture of troubled, isolated young men that matches the picture of tens of thousands of other young men who will never do this" (Grady, 2013, para. 2). In other words, given these characteristics that may describe the "typical" mass-shooter, it is more likely than not that such a person possessing these characteristics will *not* be violent.

But not everyone agrees with Dr. Swanson. Dr. Rachel Yehuda from the Mount Sinai School of Medicine's Department of Psychiatry said in the same article that Aaron Alexis confessing to hearing voices and displaying uncontrollable anger were two warning signs, and if there existed a mental health team that could be called when people like Alexis displayed such behavior, such incidents could be avoided (Grady, 2013). Dr. Michael Stone, a professor of clinical psychiatry at Columbia University, suggested the Navy-Yard Shooting *should* have been prevented:

> [Alexis] had a history of violence before ... He slipped through the cracks because people gave him way more breaks than he deserved. He told the police he was hearing voices. The police are brain-dead. They have no clue. The police in this

> generation are much more lenient about letting psychotic people get away than when I was in training 50 years ago. (Grady, 2013, para. 18)

The Atlantic Wire posted an article the day after the Navy-Yard Shooting entitled "There Will Be Another Mass Shooting. This Is What the Data Tells Us About It" (Bump, 2013a). The article went so far as to attempt to predict the next mass-shooting; using the data collected by *Mother Jones*, Bump concluded,

> The next mass shooting will take place on February 12, 2014, in Spokane, Washington. It will be committed by an emotionally disturbed, 38 year-old white man who will kill seven people and wound six more at a place he used to work using a semi-automatic handgun he purchased legally in the state. (para. 1)

The next day, Bump attempted to reassure the citizens of Spokane and surrounding areas, saying that he did not intend to scare them but rather the intention was to frighten everybody. As he stated,

> Anyone with a passing knowledge of the recent history of gun violence in America should know by now that there will be another mass killing, somewhere, and soon. Residents of Spokane need not be any more alarmed at the prospect of being gunned down while shopping or eating at a restaurant or at their school or workplace than any other American. (Bump, 2013b, para. 1)

On September 29, 2013, CBS premiered its 46th season of *60 Minutes*; the second of its three segments was entitled "Untreated Mental Illness an Imminent Danger" (Kroft, 2013). Correspondent Steve Kroft reported on the Navy-Yard Shooting that had occurred less than two weeks prior, noting it was the twenty-third mass shooting incident since 2008. In the segment, Mr. Kroft interviewed Dr. E. Fuller Torrey, a well-known psychiatrist specializing in schizophrenia and executive director at the Stanley Medical Research Institute and founder of the Treatment Advocacy Center. When Mr. Kroft asked Dr. Torrey about a possible connection between mass shootings and mental health, Dr. Torrey responded,

"About half of these mass killings are being done by people with severe mental illness, mostly schizophrenia. And if they were being treated, they would've been preventable." Mr. Kroft then postulated that had Aaron Alexis been taken to a psychiatric ward, the Navy-Yard Shooting may never have occurred.

Dr. Jeffery Lieberman, a psychiatrist at Columbia University who also specializes in schizophrenia, was asked by Mr. Kroft about schizophrenia. Dr. Lieberman stated that schizophrenia can develop in early adult life following a symptom-free childhood; this prompted Mr. Kroft to ask if a "completely normal" 20 year-old who is a "solid citizen" could become a twenty-one year-old "psychotic." Dr. Lieberman quickly replied, "Absolutely." The *60 Minutes* segment went on to discuss the struggles of people with schizophrenia stating, "[I]t's estimated that half the seven million people in the country with schizophrenia and other forms of severe mental illness are not being treated at all" (Kroft, 2013). Although Mr. Kroft correctly noted that most people who have schizophrenia are not violent, the episode continually suggested otherwise. They quoted a teenager with schizophrenia: "Basically all my voices I have are just thought—just voices telling me to harm myself or harm other people or kill people." Another interviewee stated that her schizophrenic brother in an email "said that someone was going to come to my apartment with an AR-15 and hollow point bullets and spatter my brains all over my apartment."

The episode concluded with Dr. Torrey stating,

> We have a grand experiment: what happens when you don't treat people. But then you're going to have to accept ten percent of homicides being killed by untreated, mentally ill people. You're going to have to accept Tucson and Aurora. You're going to have to accept [Seung-Hui] Cho at Virginia Tech. These are the consequences when we allow people who need to be treated to go untreated. And if you are willing to do that, then that's fine. But I'm not willing to do that.

The link between violence and mental illness has long been controversial and heavily debated. Even if such a link does exist, is it sufficient enough to warrant involuntary treat-

ment in an attempt to prevent hypothesized violence? It is a reasonable question to ask; as Dr. Swanson pointed out, it is more likely that someone possessing the characteristics of a mass murderer will not embark on a shooting rampage. So the question becomes, is it even possible to predict mass murders and other violent behavior with reasonable accuracy? As the epigraph states, prediction is not easy. Niels Bohr was a Danish physicist who made numerous contributions within his field; he was awarded the 1922 Nobel Prize in physics. Another physicist and influential figure in the development of quantum mechanics was Werner Heisenberg who worked with Bohr. The Heisenberg principle, also known as the uncertainty principle, refers to the limits in precision for physical properties of objects—the more precisely the position of an object is known, the less precisely the momentum of the object is known. Bohr and Heisenberg had a clear understanding of the uncertainty associated with prediction.

Prediction in the natural sciences is widespread; some areas have shown marked success (e.g., meteorology) whereas others have much room to improve (e.g., seismology). Like the natural sciences, prediction is a common theme in the social sciences. Whether it be prediction of tomorrow's stock prices, the next election results, or graduation rate among incoming undergraduate students, prediction in the social sciences is as widespread and just as difficult as—or, arguably, more difficult than—in the natural sciences. There will always be uncertainty involved in future events; perfection is not an attainable option when it comes to prediction and this is more or less what Bohr's message implies. The notion of uncertainty is absent from many of the quotes in the previous paragraphs; this is disconcerting. Prediction in the social sciences tends to affect the lives of many people; it goes without saying that predictions of future behavior should not be made lightly and proper methods for evaluating prediction techniques are necessary. If violent behavior is to be predicted, it must be done with the utmost ethical considerations for everyone involved.

What follows are important considerations and concepts for constructing predictive and diagnostic models in the social sciences. Although the primary focus is violence predic-

tion, the underlying points of emphases can be generalized across many types of prediction paradigms.

1.1 Statistical Toolbox: Definitions and Methods

It is necessary to introduce some terminology and techniques that will be used throughout. First, important frequencies and probabilities will be defined; this is followed by several important statistical techniques used throughout the chapters.

1.1.1 Frequencies and Probabilities

Suppose a diagnostic test is designed to determine whether a person has "it," whatever "it" may be. For example, we may be interested in predicting future violence; our test then indicates whether the person will be violent in the future. Let B denote the event that the test is positive indicating the person has "it," and $\bar{B}$, the event that the test is negative indicating that the person does not have "it." Now, consider the events of whether a person truly has "it" or truly does not have "it" and denote these two events as A and $\bar{A}$, respectively. The events B and $\bar{B}$ will be called the *diagnostic test results* and the events A and $\bar{A}$, the *states of nature*.

Given the diagnostic test result and state of nature, a 2×2 contingency table can be constructed, as shown in Table 1.1. This table provides the frequencies of marginal events (e.g., n_B is the number of people who test positive), or of joint events (e.g., n_{BA} is the number of people who have "it" and tested positive). In terms of violence prediction, n_B is the number predicted to be violent and n_{BA} is the number predicted to be and who are violent. The frequencies within the table have familiar names worth noting: n_{BA} is the number of *true positives*, $n_{B\bar{A}}$ is the number of *false positives*, $n_{\bar{B}A}$ is the number of *false negatives*, and $n_{\bar{B}\bar{A}}$ is the number of *true negatives*. Of particular importance are the marginal frequencies n_A, representing the *base frequency* for those who have "it," and $n_{\bar{A}}$,

the base frequency for those who do not. In addition, we may be interested in n_B and $n_{\bar{B}}$, the base frequencies for positive and negative diagnostic test outcomes, respectively; these are often called *selection frequencies*.

		State of Nature		
		A (positive)	$\bar{A}$ (negative)	Totals
Diagnostic	B (positive)	n_{BA}	$n_{B\bar{A}}$	n_B
Test Result	$\bar{B}$ (negative)	$n_{\bar{B}A}$	$n_{\bar{B}\bar{A}}$	$n_{\bar{B}}$
	Totals	n_A	$n_{\bar{A}}$	n

Table 1.1: A general 2×2 contingency table.

In addition to frequencies, various marginal, joint, and conditional probabilities can be defined. For example, $P(A) = {}^{n_A}/_{n}$; $P(A \cap B) = {}^{n_{BA}}/_{n}$; $P(A|B) = {}^{n_{BA}}/_{n_B}$; $P(B|A) = {}^{n_{BA}}/_{n_A}$; and so forth. These conditional probabilities are of general interest, and again it is worth noting their names. Conditionalizing on the state of nature, we have the following: $P(B|A) = {}^{n_{BA}}/_{n_A}$ is the *sensitivity* or *true positive rate* (TPR) or *recall*; $P(B|\bar{A}) = {}^{n_{B\bar{A}}}/_{n_{\bar{A}}}$ is the *false positive rate* (FPR); $P(\bar{B}|A) = {}^{n_{\bar{B}A}}/_{n_A}$ ($= 1 -$ sensitivity) is the *false negative rate* (FNR); and $P(\bar{B}|\bar{A}) = {}^{n_{\bar{B}\bar{A}}}/_{n_{\bar{A}}}$ ($= 1 -$false positive rate) is the *specificity* or *true negative rate* (TNR). Conditionalizing on the diagnostic test result, $P(A|B) = {}^{n_{BA}}/_{n_B}$ is called the *positive predictive value* or *precision*; $P(\bar{A}|\bar{B}) = {}^{n_{\bar{B}\bar{A}}}/_{n_{\bar{B}}}$ is the *negative predictive value* (NPV). The marginal probabilities represent the *base rates* for those who have "it" ($P(A)$) and those who do not ($P(\bar{A})$) (also called *prior probabilities*); those who are predicted to have "it" ($P(B)$) and those who are not ($P(\bar{B})$) (also called *selection ratios*).

It is important to note the dependency of frequencies (and, consequently, probabilities). For instance, if we know the base and selection frequencies, then the distribution of joint frequencies are subject to a single degree of freedom. As another example, given n_A and n_{BA}, $n_{\bar{B}A}$ is not free to vary. Similarly, given $P(B|A)$, the laws of probability determine that $P(\bar{B}|A) = 1 - P(B|A)$. In other words, given the true and false positive rates, the true and false negative rates are superfluous.

An Urn Model Approach.

To assist in the discussion of frequencies and probabilities, we can imagine two urns containing a total of n marbles; n_A marbles are red and $n_{\bar{A}}$ are blue. Of the n_B marbles from the first urn, n_{BA} are red and $n_{B\bar{A}}$ are blue; of the $n_{\bar{B}}$ marbles in the second urn, $n_{\bar{B}A}$ are red and $n_{\bar{B}\bar{A}}$ are blue. These frequencies can be used to construct the 2×2 contingency table in Table 1.1.

If we randomly select a marble from the set of n marbles, with replacement, various "population" probabilities (or parameters) can be modeled by such a process. For instance, the probability that we select a red marble from the set of n marbles (with replacement) is $P(A)$; the probability that we select a blue marble, given that we are selecting from the second urn, is $P(\bar{A}|\bar{B})$. The urn model can be explicitly operationalized using bootstrap sampling methods.

Confidence Intervals.

If desired, confidence intervals can be constructed for the various "true" binomial distribution parameters definable for the 2×2 contingency table. There are numerous ways of calculating these intervals (e.g., normal approximation; Clopper-Pearson, Clopper & Pearson, 1934; Agresti-Coull, Agresti & Coull, 1998; Wilson score, Wilson, 1927); depending on the estimate, some may be more desirable than others. A demonstration is provided using the most common method based on a normal approximation.

To construct a confidence interval estimate various conditional probabilities are assumed to characterize an underlying binomial distribution. For instance, let $\hat{p}_{B|A} \equiv P(B|A)$ be the estimated sensitivity of a test. Using the normal approximation to the binomial distribution, a $(1-\alpha) \times 100\%$ confidence interval for the true sensitivity, $p_{B|A}$, follows:

$$\hat{p}_{B|A} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_{B|A}(1-\hat{p}_{B|A})}{n_A}},$$

where for the cumulative standard normal distribution function, $\Phi(\cdot)$, we have $\Phi(z_{1-\alpha/2}) = P(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2$. Similarly, a confidence interval for the true positive predictive value, $p_{A|B} \equiv P(A|B)$, is given as

$$\hat{p}_{A|B} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_{A|B}(1-\hat{p}_{A|B})}{n_B}}.$$

For further discussion and demonstration using confidence intervals for conditional probabilities, the reader is referred to Pepe (2003, pp. 22–23). Alternatively, bootstrap methods can be used to construct confidence intervals (e.g., Jhun & Jeong, 2000).

Bayes' Theorem.

From Table 1.1, it is readily seen that $P(A|B) = P(A \cap B)/P(B)$; or equivalently, $P(A \cap B) = P(A|B)P(B)$. Similarly, $P(B|A) = P(B \cap A)/P(A) \Leftrightarrow P(B \cap A) = P(B|A)P(A)$. Because $P(A \cap B) = P(B \cap A)$, we have the following equality: $P(A|B)P(B) = P(B|A)P(A)$; solving for the first conditional probability gives a relationship between conditional probabilities:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

This relationship between conditional probabilities is known as *Bayes' Theorem*; it is one of the most important—and useful—results in probability theory.

Because $P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$, Bayes' Theorem can be rewritten as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}.$$

In words, Bayes' Theorem shows that the positive predictive value can be determined when we know the sensitivity, specificity, and base rate. Similar relationships exist for the negative predictive value and the sensitivity and specificity.

Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, and Woloshin (2007) suggest the use

of natural frequencies over probabilities for conveying information. For example, if the positive predictive value is equal to .40, we would say that for every ten positive tests four people truly have "it." Gigerenzer et al. (2007) noted the ease of calculating conditional probabilities using natural frequencies and Bayes' Theorem, and showed that training physicians to translate conditional probabilities to frequencies alleviates much of the confusion associated with conditional probabilities.

Using an example similar to that provided by Gigerenzer et al. (2007), but now stated in terms of five-year recidivism among sexual offenders, suppose one is told the following facts: the probability a sexual offender will recidivate is .10; if a sexual offender will recidivate, the probability he will be so predicted is .38; if the sexual offender will not recidivate, the probability he is so predicted is .12. If asked to state the probability that a sexual offender will recidivate given that he is so predicted, the answer can be found using Bayes' Theorem; Gigerenzer et al. (2007) show that when stated in terms of frequencies, this problem is more likely to be correctly answered. For example, if the information is provided as follows: out of 1000 sexual offenders, 100 will recidivate; of the 100 that will recidivate, 38 will be so predicted; and out of the 900 that will not recidivate, 108 will be so predicted. The correct answer using probabilities is

$$\frac{(.38)(.10)}{((.38)(.10)+(.12)(.90))} = .26;$$

for natural frequencies:

$$\frac{38}{108+38} = .26.$$

1.1.2 Statistical Discrimination and Classification

Suppose one wishes to distinguish between two or more groups of observations or individuals using a set of variables. Statistical discrimination methods usually find linear combinations of the variables of interest that characterize and distinguish the groups. As

an example, suppose we have two populations, π_0 and π_1; they could represent, respectively, individuals who will not be violent in the future and individuals who will. For simplicity, consider only one variable for classifying individuals and assume these individuals come from one of the two populations. For convenience in the initial presentation, the distributions of the two populations are assumed to follow normal distributions with the same variance, σ^2, but with different means, μ_0 and μ_1 (without loss of generality, let $\mu_0 \leq \mu_1$). The two normal densities are denoted by $f_0 \equiv f_0(X; \mu_0, \sigma^2)$ and $f_1 \equiv f_1(X; \mu_1, \sigma^2)$, where X is a random variable (e.g., a randomly selected individual from the respective population).

Given an observation (or individual), say x, we wish to "classify" the observation into one of the two populations. For example, our concern may be whether that individual will be violent in the future. The true state of nature for the individual is whether he or she will be violent; the decision is whether we classify the individual as someone who will be violent. This decision is made based on a cutscore or criterion point, denoted x_c, for the variable of interest. If $x \leq x_c$ we allocate the individual to π_0; if $x > x_c$, x is allocated to π_1. Defining a decision function with respect to x as $d \equiv d(x)$; $d = 0$ if we allocate x to π_0 and $d = 1$ if we allocate x to π_1. For example, the variable might be an individual's Psychopathy Checklist (PCL) score; if the PCL score is above a specified threshold, x_c, he or she would be classified as coming from the population committing an act of violence in the future. If that individual has a PCL score less than or equal to x_c, he or she will be classified as coming from the population not committing an act of violence. Given an observation, we again have a 2×2 table, this time representing the error at the individual level as shown in Table 1.2; here, α and β represent, respectively, the false positive and false negative probabilities. Figure 1.1 gives a visual representation of the two distributions with the error probabilities highlighted.

Ideally, we would like to minimize both α and β; but because the two distributions overlap, these errors are intertwined and we cannot decrease one without increasing the other. Instead, we focus on choosing a cutscore x_c such that $\alpha + \beta$ is minimized. Figure 1.2 displays this situation; $x_{c'}$ represents the cutscore where $\alpha + \beta$ is minimized. It should be

		State of Nature	
		f_1	f_0
Decision	f_1	$1-\beta$	α
	f_0	β	$1-\alpha$

Table 1.2: A 2×2 table representing the error rates for a decision.

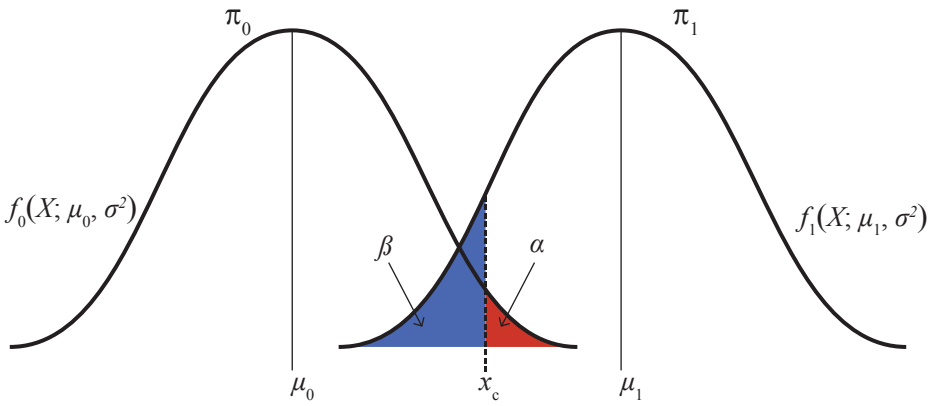

Figure 1.1: Two normal distributions representing two populations. For a given cutscore, x_c, two types of error probabilities are present: α represents the probability of a false positive, and β, the probability of a false negative.

clear that $x_{c'}$ is the point on the horizontal axis at which the two distributions intersect; that is, where $f_0 = f_1$ (or, equivalently, where $f_0/f_1 = 1$). Therefore, our decision function can be defined as

$$d(x) = \begin{cases} 1 & \text{if } \frac{f_0}{f_1} > 1, \\ 0 & \text{if } \frac{f_0}{f_1} \leq 1. \end{cases}$$

Generalizing, we can also assign prior probabilities to each of the population distributions, denoted p_0 and p_1, where $p_0 + p_1 = 1$; p_0 is the probability that a given observation, x, comes from the population π_0 and p_1 is the probability that the observation comes from the population π_1. In terms of nonviolent and violent populations and using previously discussed definitions, p_0 is the base rate for nonviolence (i.e., $P(\bar{A})$); p_1 is the base rate for violence (i.e., $P(A)$). To account for differing prior (base rate) probabilities, we want to chose an x_c to minimize the *total probability of misclassification*, (i.e., $p_0\alpha + p_1\beta$). Our decision function

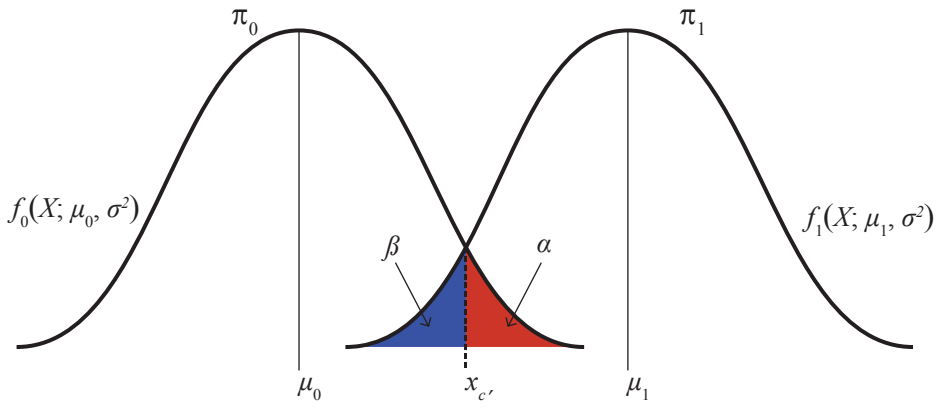

Figure 1.2: Cutscore minimizing $\alpha + \beta$, denoted $x_{c'}$, for classifying an unknown observation into one of two population distributions.

now becomes

$$d(x) = \begin{cases} 1 & \text{if } \frac{f_0}{f_1} > \frac{p_1}{p_0}, \\ 0 & \text{if } \frac{f_0}{f_1} \leq \frac{p_1}{p_0}. \end{cases}$$

Finally, a cost function, $C_k(j)$, can be included that assigns a cost to allocating x to population π_k when x actually comes from population π_j; $C_1(0)$ is the cost for allocating x to population π_1 when x really comes from population π_0 and $C_0(1)$ is the cost of allocating x to population π_0 when x truly comes from population π_1. We now choose x_c to minimize the expected cost of misclassification, $C_0(1)p_0\alpha + C_1(0)p_1\beta$. Our updated decision function becomes

$$d(x) = \begin{cases} 1 & \text{if } \frac{f_0}{f_1} > \left(\frac{C_0(1)}{C_1(0)}\right)\left(\frac{p_1}{p_0}\right), \\ 0 & \text{if } \frac{f_0}{f_1} \leq \left(\frac{C_0(1)}{C_1(0)}\right)\left(\frac{p_1}{p_0}\right). \end{cases}$$

For the assumed normal distributions in this discussion and because

$$f_k(x) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\left((x-\mu_k)^2\left(2\sigma^2\right)^{-1}\right)\right\},$$

the ratio $\frac{f_0}{f_1}$ equals $\exp\left\{\left((x-\mu_1)^2 - (x-\mu_0)^2\right)\left(2\sigma^2\right)^{-1}\right\}$. Taking the natural logarithm of

the above equation,

$$\begin{aligned}\ln\left(\frac{f_0}{f_1}\right) &= \frac{(x-\mu_1)^2-(x-\mu_0)^2}{2\sigma^2} \\ &= \frac{x^2-2x\mu_1+\mu_1^2-x^2+2x\mu_0-\mu_0^2}{2\sigma^2} \\ &= \frac{2x(\mu_0-\mu_1)-(\mu_0-\mu_1)(\mu_0+\mu_1)}{2\sigma^2}.\end{aligned}$$

Thus, using our criterion for the decision function based on minimizing the expected cost of misclassification and taking the logarithm of both sides, $d = 1$ when

$$\begin{aligned}\ln\left(\frac{f_0}{f_1}\right) &> \ln\left[\left(\frac{C_0(1)}{C_1(0)}\right)\left(\frac{p_1}{p_0}\right)\right] \\ \frac{2x(\mu_0-\mu_1)-(\mu_0-\mu_1)(\mu_0+\mu_1)}{2\sigma^2} &> \ln\left[\left(\frac{C_0(1)}{C_1(0)}\right)\left(\frac{p_1}{p_0}\right)\right] \\ x &> \frac{1}{2}(\mu_0+\mu_1)+\ln\left[\left(\frac{C_0(1)}{C_1(0)}\right)\left(\frac{p_1}{p_0}\right)\right]\left(\frac{\sigma^2}{\mu_0-\mu_1}\right) \equiv x_c.\end{aligned}$$

Because μ_0, μ_1, σ, p_0, and p_1 are assumed fixed, the cutscore, x_c, is dependent only upon the cost functions $C_0(1)$ and $C_1(0)$. If $C_0(1) = C_1(0)$, the cutscore is determined strictly from the fixed parameters. Note that if we have a desired cutscore in mind, the cost ratio can be determined given this cutscore.

The classification models can be extended to a multivariate framework. The population π_i $(i = 0, 1)$ is characterized by an $m \times 1$ vector of random variables, $\mathbf{x} \sim \mathcal{N}_m(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. For example, $\mathbf{x}$ could be a 4×1 vector representing patients' age, IQ level, number of previous arrests, and PCL score.

Although we assumed the populations follow normal distributions with equal variances, these assumptions can be relaxed. The normal assumption can be relaxed to one of a unimodal distribution. If variances are not assumed equal, a quadratic classifier can be used.

1.1.3 Signal Detection Theory

Signal Detection Theory (SDT) is a framework for assessing uncertainty in decision making; it is also commonly used in evaluating diagnostic measures in psychology and medicine, among other areas. SDT allows the researcher to quantify a diagnostic tool's ability to distinguish meaningful patterns and known processes (called the signal) from random patterns or chance processes (called the noise). Based on the given data, three important parameters are estimated: the strength of the signal relative to the noise, called the *discriminability index* (denoted d'); the decision criterion of the diagnostic measure, called the *cutscore* (denoted x_c); and an indication of *response bias* (denoted β).

Given that the signal is truly present, a correct decision (i.e., stating that the signal is present) is called a *hit* (or true positive); an incorrect decision (i.e., stating that noise is present) is called a *miss* (or false negative). When noise is truly present, a correct decision (i.e., stating that noise is present) is called a *correct rejection* (or true negative); an incorrect decision (i.e., stating that the signal is present) is called a *false alarm* (or false positive). Similar to our earlier discussion, these can be framed in terms of rates: the *hit rate* (true positive rate; sensitivity) is the number of correct responses given the presence of the signal; the *false alarm rate* (false positive rate; $1 -$ specificity) is the number of incorrect responses given the presence of noise. In SDT, these are the two rates of interest; the other two are easily calculated from them and therefore provide no additional information.

Two distributions can be constructed; one represents the signal, the second represents the noise (see Figure 1.3). To simplify our discussion, they are considered to be normal distributions with equal variances but differing means. Without loss of generality, the mean of the signal distribution is assumed greater than that of the noise distribution. The difference between the two means of the distributions is indicative of the strength of the signal in relation to the noise; this is the discriminability index, d'. Based on the hit and false alarm rates, d' is calculated as $\Phi^{-1}(\text{hit rate}) - \Phi^{-1}(\text{false alarm rate})$, where $\Phi(\cdot)$ represents

the cumulative standard normal distribution. We note that d' requires several parametric assumptions regarding normal distributions; there are nonparametric alternatives, but these are not discussed here (e.g., the measure commonly denoted as A'; Pollack & Norman, 1964). A d' of 0 indicates that the diagnostic measure cannot distinguish signal from noise. The criterion point distinguishing a positive response—with respect to the presence of the signal—from a negative response is the cutscore, x_c. Given x_c, the area under the noise distribution to the right of x_c represents the probability of a false alarm (i.e., the false alarm rate); the area under the signal distribution to the left of x_c represents the probability of a miss (i.e., $1 -$ hit rate). Another detail gathered from x_c is the density (or height) of the two distributions corresponding to x_c. These two densities represent the likelihoods of signal (the height of the signal distribution at x_c) and noise (the height of the noise distribution at x_c). The response bias of the test, β, is the ratio of these two: $^{\text{signal likelihood}}/_{\text{noise likelihood}}$. If $\beta > 1$, the diagnostic test favors (i.e., is biased toward) a positive decision (stating that the signal is present) over a negative one (stating that the signal is absent); if $\beta < 1$, the diagnostic test favors a negative decision over a positive one; if $\beta = 1$, there is no response bias as neither decision is favored over the other. Another way of thinking about β is that when an observation x is greater than x_c, then the ratio of the two likelihoods at the point x is greater than β.

More realistically, the two distributions can differ not only in their means but in variances as well. Furthermore, the prior probability of the noise distribution is not likely to be equal to the prior for the signal distribution.

ROC Plots.

There are numerous (possibly infinitely many) different choices for a cutscore, x_c. At each cutscore, we have a unique hit and false alarm rate and this can be plotted in a two-dimensional plot with the hit rate (often labeled sensitivity or true positive rate) along the vertical axis and false alarm rate (often labeled $1-$specificity or false positive rate) along the

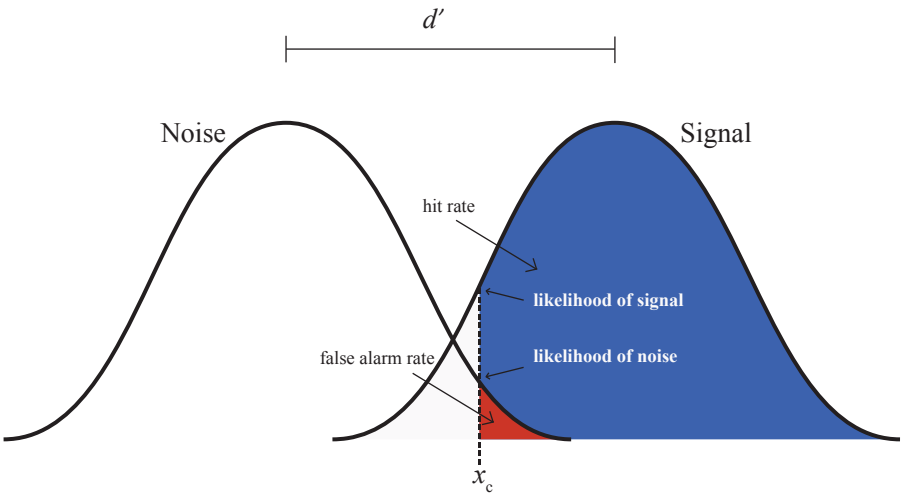

Figure 1.3: Two normal distributions representing the distribution of the signal (right) and the noise (left).

horizontal axis. This plot is referred to as a *receiver operating characteristic* (ROC) curve (see Figure 1.4).

The minimum along each axis is 0; the maximum is 1. The point $(0, 0)$ corresponds to hit and false alarm rates of 0; that is, a cutscore larger than any attainable score. This is equivalent to stating that the signal is always absent. The other extreme is at $(1, 1)$ and corresponds to hit and false alarm rates of 1; that is, it is representative of a cutscore that is smaller than any attainable score. In this situation the signal is always stated to be present. Neither of the two extreme situations are useful—a diagnostic test is unnecessary if we are consistently predicting one way or the other. What is of interest are the values in between that correspond to meaningful cutscores. By moving along the curve one can choose a cutscore corresponding to different hit and false alarm rates.

A 45° line is often placed on the plot, running from the point $(0, 0)$ to $(1, 1)$. This line is called the *line of no discrimination* and represents a diagnostic test performing no better than chance; ideally all points on the ROC curve are above this line. The discriminability index, d', is represented by the distance of the curve to the line of discrimination; specifically, it is the orthogonal distance measured at the point $(.5, .5)$ to the curve. As the

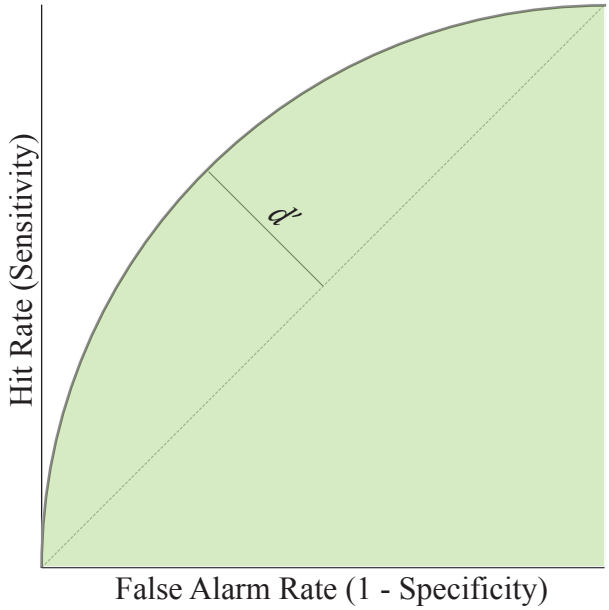

Figure 1.4: A receiver operating characteristic (ROC) plot. The points along the curve represent the different hit and false positive rate combinations achieved using different cutscores. The shaded green area represents the area under the curve.

two distributions are further spread apart, this distance becomes larger.

As a measure of diagnostic accuracy, the *area under the ROC curve* (AUC; also called the *concordance index*) is commonly used. The total area ranges from 0 to 1; however, an AUC less than .50 is infrequently found because it represents a test that performs worse than chance. The AUC can be interpreted as the probability that a randomly selected individual who has "it" will have a larger score on the diagnostic test than a randomly selected individual who does not have "it." There are ways to test whether two ROC curves are different from each other (e.g., whether one diagnostic test outperforms another), or when an ROC curve is different from a diagnostic test that is no better than chance (i.e., an ROC curve equal to the line of no discrimination). A final point is that the ROC curve is independent of the base rates; this will be discussed further in Chapter 4.

1.1.4 Decision Trees

Decision trees, commonly referred to as *Classification and Regression Trees* (CART; Breiman, Friedman, Olshen, & Stone, 1984), are popular machine learning techniques generally used for prediction; decision trees have been constructed by several researchers as an actuarial tool for predicting violent and dangerous behavior (e.g., see Steadman et al., 2000).

Consider an $n \times p$ data matrix, $\mathbf{X}$, containing n observations measured across p predictor variables, and an $n \times 1$ vector $\mathbf{y}$ containing n observations measured across a single outcome variable, Y. The outcome variable contained in $\mathbf{y}$ is the variable of interest with respect to prediction. If the outcome variable is continuous, regression trees are constructed; if the variable is categorical, classification trees are constructed. In violence prediction, the outcome variable is typically binary (i.e., categorical) representing the presence or absence of an act of violence; because of this specification, our focus is solely on classification trees.

Classification trees are constructed by first splitting the data into two disjoint subsets based on one of the p predictor variables. Within each subset, further partitioning of the data is done, and within the resulting subsets this process continues until some user-specified stopping criterion is reached; the complete procedure is known as *recursive partitioning*. An observation that falls into a subset with no further splits (called a *terminal node*) is classified based on all the observations within that subset; typically this is the modal observation (i.e., the most prevalent outcome within the node).

The first split occurs at the *root node* of the tree; extending *branches* from the root node lead to subsample nodes, called *leafs*. As mentioned, the splits continue until a specified criterion is met, such as constraining the minimum number of observations in a given leaf or based on significance testing (e.g., AID; see below). After a tree is created, it can be *pruned* to reduce the number of branches, eliminating those that add less to the tree's predictive ability.

Given K classes for the outcome variable Y, we can split the data into two groups

based on the classes. The total number of possible splits is $2^{K-1} - 1$. For a binary outcome, this is equal to one unique split of the data. In regards to predicting violent behavior, there are those who were violent and those who were not. A perfect split based on the p predictors classifies all those who are violent as violent and those who are not violent as nonviolent. The goal is to come as close to a perfect split as possible.

There are several ways to measure a "best" fitting tree, one being the *Gini index* (Gini, 1912). Given K classes, the Gini index for the two groups is defined by the proportion of the group coming from the K classes. Let r_k, $k = 1, \ldots, K$, represent the proportion in the group from the kth class. The Gini index is

$$G = \sum_{k=1}^{K} r_k(1 - r_k) = 1 - \sum_{k=1}^{K} r_k^2.$$

The Gini index can be thought of as a measure of impurity. Note that if $r_k = 1$, $r_{k'} = 0$ for $k' \neq k$ and consequently, $G = 0$. If $r_1 = \cdots = r_K = 1/K$, the index is maximal at $G = 1 - 1/K$. Because there are only two groups, the Gini index is the same for both groups.

Decision trees are popular because they are easy to interpret, but they are not the most powerful machine learning method in terms of predictive accuracy. Predictive accuracy can be enhanced by ensemble methods such as *tree bagging*, *tree boosting*, and *random subspace methods* (i.e., *random forests*). Random forests are discussed in detail in Chapter 5.

Automatic Interaction Detection (AID).

Morgan and Sonquist (1963) proposed an algorithm designed to partition a sample into disjoint subsamples where the means within the subsample groups explain more of the variation than any other subsample. This method is known as *automatic interaction detection*, or AID, and can be used to construct decision trees; at each level an optimal split is found for a given variable among samples/subsamples in terms of variance-accounted-for. If the outcome is binary, this involves splitting the sample into two subsamples such that

each subsample is optimal in terms of misclassification. The splitting continues within each subsample until an optimal split is not possible or a constraint is no longer met (e.g., a minimum number of observations in a subsample).

One of the more popular AID techniques is the Chi-Squared AID, or CHAID (Kass, 1980). The CHAID algorithm selects predictor variables that yield the most significant split, typically determined by a Bonferonni-adjusted p-value obtained using Pearson's chi-squared test (Pearson, 1900b), assuming the outcome variable is categorical. This is carried out until no significant splits remain, using an a priori threshold set by the user (typically $p < .05$).

1.2 Actuarial Tools

There are two actuarial tools given considerable focus in the subsequent chapters. One is the Classification of Violence Risk (COVR) to assess dangerousness; the second is the Static-99 and Static-2002 for predicting sexual and violent recidivism.

1.2.1 The Classification of Violence Risk (COVR)

The Classification of Violence Risk (COVR; Monahan et al., 2006) is an assessment tool developed from the MacArthur Violence Risk Assessment Study (VRAS). The COVR is a computer-implemented instrument designed to estimate the risk of violence in psychiatric patients; given the appropriate credentials, it is available for purchase from Psychological Assessment Resources (PAR; `http://www4.parinc.com`). The COVR assigns patients into five risk groups defined by the "likelihood that the patient will commit a violent act toward another person in the next several months" (Monahan et al., 2006, p. 728). Table 1.3 gives the five risk groups defined by their best point risk estimates and their 95% confidence intervals.

The development of the COVR is detailed in Monahan et al. (2001; also see Steadman et al., 2000, Monahan et al., 2000, and Banks et al., 2004 for less detailed reviews). The

Category	Risk	Point Estimate	95% CI
5	Very High	.76	[.65, .86]
4	High	.56	[.46, .65]
3	Average	.26	[.20, .32]
2	Low	.08	[.05, .11]
1	Very Low	.01	[.00, .02]

Table 1.3: The five risk categories for the Classification of Violence Risk (COVR) diagnostic test along with point estimate risks (in probabilities) and respective confidence intervals (CI) (Monahan et al., 2006).

COVR was based on a sample of 939 recently-discharged patients from acute inpatient psychiatric facilities in three locations within the United States: Pittsburgh, Pennsylvania; Kansas City, Missouri; and Worcester, Massachusetts. Patients were restricted to those who were white, African-American, or Hispanic; English-speaking; between the ages of 18–40; and charted as having thought, personality, or affective disorder, or engaged in substance abuse.

According to the original MacArthur study (Monahan et al., 2001), violence is defined as "acts of battery that resulted in physical injury; sexual assaults; assaultive acts that involved the use of a weapon; or threats made with a weapon" (p. 17). A second category of violent incidents was labeled as "other aggressive acts" (Monahan et al., 2001, p. 17) including non-injurious battery; verbal threats were not considered. The outcome variable of violence is a dichotomous variable—either the patient committed an act of violence or did not. It does not consider the number of violent acts or their severity. The patients were interviewed once or twice during the twenty weeks after discharge. Of the 939 patients, 176 were considered violent; thus, the base rate for violence in this sample is 18.7 percent.

The authors identified 134 potential risk factors, listed in detail in Monahan et al. (2001, Appendix C). Using SPSS's CHAID algorithm (SPSS, 1993), the authors developed a classification tree based on the given risk factors. The final classification model used an iterative classification tree (ICT) approach; after an initial classification tree was developed,

those who were still unclassified (i.e., those within .09 to .37 estimated probabilities of committing violence according to the model) were reanalyzed using the CHAID algorithm. After four iterations, 462 patients were classified as low risk (less than .09 probability of committing violence), 257 were classified as high risk (greater than .37 probability of committing violence), and 220 remained unclassified. The cutoffs of .09 and .37 were chosen because they represent, respectively, one half and twice the base rate of violence in the sample.

The authors' goal was to create an actuarial tool that was "clinically feasible"; thus, it included only those risk factors that could easily be computed, eliminating 28 of the original 134 risk factors that "would be the most difficult to obtain in clinical practice" (Monahan et al., 2001, p. 108) as determined by the length of the instrument measuring the risk factor (more than twelve items was considered too long) or if the risk factor was not readily or easily available to mental health professionals. After doing so, the same ICT method was applied to the 106 remaining risk factors using three iterations.

The correlation between the predictions made by the clinically-feasible and original ICT models was .52. The authors noted the low correlation:

> The fact that these [two] prediction models are comparably associated with the criterion measure, violence (as indicated by the ROC analysis), but only modestly associated with each other [as indicated by the correlation coefficient], suggested to us that each model taps into an important, but different, interactive process that relates to violence. (p. 117)

The authors then constructed nine additional ICT models using the 106 clinically-feasible variables; for each of the nine trees the authors "forced a different initial variable" (p. 118; i.e., the root node for each of the ten trees was distinct). The ten models led to ten classifications for each individual (high, average, or low) and each individual was assigned a score corresponding to their risk level (1, 0, or -1, respectively); the scores were then summed to create a composite score ranging from -10 to 10. The authors remarked, "As two models predict violence better than one, so ten models predict violence better than two

(i.e., the area under the ROC curve was .88 for ten models compared to .83 for two models)" (p. 122).

The authors questioned whether ten models were necessary; to determine this empirically they performed stepwise logistic regression and concluded that only five of the ten were needed, leading to composite scores ranging from -5 to 5 (the area under the ROC curve remained the same). The composite scores were divided into five distinct groups based on the following ranges: $[-5, -3]$, $[-2, -1]$, $[0, 1]$, $[2, 3]$, and $[4, 5]$ (these five groups correspond to, respectively, the very-low, low, average, high, and very-high risk groups found in Table 1.3; the probabilities represent the proportion of those violent within each group). The authors did not cross-validate their model; they did, however, use bootstrapping to estimate the confidence intervals provided in Table 1.3.

1.2.2 Static-99 & Static-2002

The Static-99 (Hanson & Thornton, 2000) is an actuarial assessment tool for estimating the long-term risk of recidivism among male sexual offenders over the age of 18. According to its website (`www.static99.org`), the Static-99 "is the most widely used sex offender risk assessment in the world" (Static-99, 2013). The Static-99 is comprised of two other risk assessment measures: the Rapid Risk Assessment for Sex Offense Recidivism (RRASOR; Hanson, 1997) and Thornton's Structured Anchored Clinical Judgement scale (SAC-J; Grubin, 1998). The Static-99 consists of ten items, nine items are dichotomous, one is polytomous; the items are summarized in Appendix B. The scores across the ten items are summed, with the patient falling into one of four categories based on total score: high (6+), medium-high (4–5), medium-low (2–3), or low (0–1). The authors used survival analysis to estimate the recidivism rates after 5, 10, and 15 years.

In 2003, Hanson and Karl developed the Static-2002, a revised version of the Static-99 developed for "increased coherence and conceptual clarity [and to] ... improve the consistency of the scoring criteria" (p. 1). According to the authors, the Static-2002 has a "level of

predictive accuracy ... similar [to the Static-99]" and "predicted any violent recidivism as well as it predicted sexual recidivism, and predicted violent recidivism better than Static-99" (Hanson & Karl, 2003, p. 17). Note that predictive accuracies are being measured by ROC values (see Chapter 4). The coding for the Static-2002 can be found in Appendix B; the maximum score is 14.

1.3 Road Map

1.3.1 Chapter 2: A History of Violence Prediction

Violence is wide-spread in American culture; it is a common theme in popular television programming, movies, music, video games, and various media outlets. Dangerous behavior and violent acts are usually the "newsworthy" stories and often take precedence over less exciting but arguably more important topics. Because violence plays such a significant role in our society, it is natural to want to predict dangerous behavior. Chapter 2 explores the history of violence prediction, particularly focusing on the use of actuarial methods to do so. Statistical methods are commonly used in the judicial system and have been developed to predict parole violations, juvenile delinquency, future dangerous behavior, violent and sexual recidivism, and more recently, future crime in general.

1.3.2 Chapter 3: It's All About the Base Rates

Some fifty years ago, Meehl and Rosen (1955) stated a condition under which a diagnostic test would be "clinically efficient." As they defined it, clinical efficiency refers to a situation where prediction by a diagnostic test is better than prediction using only the raw base rates. Although cited extensively, the actual importance of the Meehl and Rosen condition for deciding on when to use a diagnostic instrument seems generally ignored in the literature. Chapter 3 reviews the Meehl and Rosen condition and offers two equivalent

others attributed to Dawes (1962) and Bokhari and Hubert (introduced in the chapter). The relationships are developed between those various equivalent conditions for clinical efficiency and several measures of association in a 2×2 contingency table (e.g., the Goodman-Kruskal lambda coefficient, the odds ratio, and relative risk). Data collected in attempts to predict dangerous behavior, such as the Classification of Violence Risk (Monahan et al., 2001), are provided as illustrative examples.

1.3.3 Chapter 4: Hiding Behind the AUC

The area under the receiver operating characteristic (ROC) curve (AUC) is a widely-used measure for assessing diagnostic test performance. Despite this popularity, the AUC is not a good measure of a test's overall performance when applied to populations having differing base rates for the characteristic being assessed. Chapter 4 presents several reasons and accompanying illustrative examples in violence prediction as to why the AUC measure is subject to bias and is a misleading indicator of clinical efficiency. As an alternative measure of a test's performance, we suggest the use of the positive and negative predictive values; these incorporate base rate information and indicate whether a test will actually outperform base-rate prediction.

1.3.4 Chapter 5: Lack of Cross-Validation

Cross-validation is an important aspect in building predictive models but it is commonly ignored. Chapter 5 discusses some general methods for cross-validation and its importance. Data from the MacArthur Violence Risk Assessment Study (Monahan et al., 2001) are used to develop logistic regression, linear discriminant, and decision tree models based on cross-validation methods. The results are compared to the non-cross-validated results presented in Monahan et al. (2001).

1.3.5 Chapter 6: The Variance-Bias Trade-off

Chapter 6 looks at the variance-bias trade-off in prediction. The complexity of predictive models determines the amount of variance associated with the model as well as its bias. The variance-bias trade-off is called as such because bias generally increases as the variance decreases. This phenomenon has major implications for the prediction in new samples, generally resulting in reduced accuracy or goodness-of-fit measures. Shrinkage estimators, such as the Kelley's True Score estimator (Kelley, 1923) or the James-Stein estimator (W. James & Stein, 1961), will reduce the error associated with prediction, but at the cost of increasing the bias.

1.3.6 Chapter 7: An Overview of Violence Prediction

Chapter 7 reviews several studies of violence prediction, and summarizes the different diagnostic tools and their failure to predict violence with clinical efficiency. The chapter examines the Classification of Violence Risk (Monahan et al., 2001) and Static-99 and Static-2002 actuarial tools. The actuarial methods fail to meet the criterion for clinical efficiency and thus we caution against their use in practice.

Chapter 2

A History of Violence Prediction

"Violence does even justice unjustly."

— Thomas Carlyle, *Past and Present*

Predicting violent and dangerous behavior has long been a part of American society, and the methods, consequences, and ethical dilemmas associated with such predictions are often cause of much debate. This chapter chronicles the history of predicting violent and dangerous behavior. Particular focus is given to predictions using actuarial methods, from Burgess expectancy tables and Glueck predictive tables to the Classification of Violence Risk and the Violence Risk Assessment Guide to police jurisdictions predicting crime before it happens.

2.1 Introduction

On December 14, 2012, tragedy struck Newtown, a small Western Connecticut town. Adam Lanza, a 20-year-old man, shot through the glass door of Sandy Hook Elementary School after earlier killing his mother (S. Gray, 2012). Upon entering the school around 9:30 a.m. local time (Bryan, 2012), he proceeded to murder 20 school children and six staff members of the school in a short period of time before eventually taking his own life (Candiotti & Aarthun, 2012). The shooting was the second deadliest in the history of the United States (behind the 32 Virginia Tech killings in 2007; Bratu, 2012).

In the previous decade, Newtown had one reported homicide (Candiotti & Aarthun, 2012); residents could not believe that such a devastating event would occur in their town

with a population just under 30,000. Yet it did, and Newtown is not the only such place to suffer from such a tragic event. As argued by Newman, Fox, Roth, Mehta, and Harding (2005), the roots of rampage school violence are embedded within the communities in which they occur. The authors define rampage shootings as shootings that "involve attacks on multiple parties, selected almost at random" (p. 15). They argue toward preventive measures aside from profiling and prediction, noting that "[t]he U.S. Secret Service determined that it is impossible to profile school shooters ... [e]ven when the cases are limited to rampage school shootings, there is still too much diversity among them to predict which students could become rampage shooters" (p. 268).

The mental health of Adam Lanza immediately came into question after the Newtown shootings. It was reported that Lanza was diagnosed with sensory integration disorder (SID) when he was a young boy and dubiously noted that "[t]here is no known link between SID and violent behavior" (Breslow, 2013). SID, which is not recognized in the current Diagnostic and Statistical Manual of Mental Disorders (DSM-V; American Psychiatric Publishing, 2013), is marked by increased or decreased stimuli sensitivity.

On the day of the Newtown shooting, a Boise, Idaho blogger named Liza Long (a.k.a., "The Anarchist Soccer Mom") wrote an article entitled *Thinking the Unthinkable* (Long, 2012). In her blog post, she discusses the difficulties of raising a mentally-ill 13 year-old son who "terrifies" her. In the emotional article, she recalls a few weeks earlier when her son threatened her with a knife and how her 7 and 9 year old children "knew the safety plan" (run to the car and lock the doors). Her son is on numerous medication plans (e.g., antipsychotics), but as Long explains, she "still [doesn't] know what's wrong with [him]." Her son displays an exceptional ability for mathematics and science and has an IQ that is "off the charts." Long states that her gifted son is in a good mood most of time, but "when he's not, watch out" and that it is "impossible to predict what will set him off." Long shares many details of her son's episodic fits of rage and expresses her desire for help. She even states "I am Adam Lanza's mother," implying that her son may be the next Adam Lanza.

This quote quickly became viral in the Internet media. Her message is simple and she states it explicitly: "it's time to talk about mental illness."

The social worker for Miss Long's son provided an option to alleviate the fear she lives in: "get [her son] charged with a crime." Long quotes her son's social worker: "If he's back in the system, they'll create a paper trail. That's the only way you're ever going to get anything done. No one will pay attention to you unless you've got charges." Torrey, Kennard, Eslinger, Lamb, and Pavle (2010) found that in the United States, there are three times more mentally-ill persons in jails and prisons than in hospitals, with some states (Arizona and Nevada) having a ratio of near ten to one. The United States Department of Justice reported in 2006 that "more than half of all prison and jail inmates had a mental health problem" (D. J. James & Glaze, 2006, p. 1). Liza Long's son's social worker and the numbers cited point to a disconcerting belief that mental illness implies criminality. Miss Long states that she does not want or feel her son should be locked away in prison; she follows this sentiment with: "our society, with its stigma on mental illness and its broken healthcare system, does not provide us with other options."

Three days after the Newtown shooting and Liza Long's blog post, the women's lifestyle website xoJane printed an anonymously-written article entitled "I am Adam Lanza's psychiatrist" (2012). The author notes that the field of psychiatry suffers from numerous limitations in predicting violence from mental illness. The author asks, given a "ticking time bomb" patient, "At what point [does the author] inform the authorities when no specific plans to commit violence are mentioned to me, but the words start to become more terror-inducing." On January 15, 2013, legislators from New York State signed into law the New York Secure Ammunition and Firearms Enforcement (SAFE) Act in an attempt to answer that question. According to the SAFE Act, "mental health professionals [are required] to report to their local director of community services or his/her designees when, in their reasonable professional judgment, one of their patients is 'likely to engage in conduct that would result in serious harm to self or others' " (New York State, 2013). In other words, the

State of New York now requires its mental health professionals to predict violence in their patients.

2.2 Clinical Prediction

In 1915, the Los Angeles commanding sergeant of the Police Juvenile Bureau, Leo Marden, conjectured a cause for crime. *The Independent* (1915) reports Marden "has made careful study of the matter of juvenile crime"; continuing,

> "By keeping an exhaustive record of such matters," says Mr. Marden, "I find that over ninety per cent of the boys under twenty-one years of age who are arrested and brought to my office are cigaret smokers. The prisoners are almost without exception stunted in growth and under-developed in mind. Therefore, cigarets *must* [emphasis added] have something to do with crime, and it is my idea to treat the cause of crime." (p. 231).

2.2.1 Capital Punishment

Predictions of dangerousness[1] by psychiatrists or other experts, henceforth referred to as clinical predictions, were, and to some extent still are, an integral part of the legal system. In the state of Texas the death penalty is imposed only after the defendant is found guilty of a crime that warrants capital punishment (for the specific "qualifications" of Texas and all other States that allow the death penalty, see `http://www.deathpenaltyinfo.org/`). The State of Texas requires that, after the jury has determined beyond a reasonable doubt that the defendant is guilty, a separate trial phase occurs where the jurors must determine whether the defendant will receive the death penalty or life in prison without parole. In this punishment phase, the jury must unanimously decide "whether there is a probability that the

[1]Megargee [1976] argues a more appropriate term instead of "dangerousness" is "dangerous behavior," as dangerousness implies the existence of an innate, rather than acquired, characteristic of an individual. Although we agree, we will generally use the more common, albeit less appropriate, term dangerousness.

defendant would commit criminal acts of violence that would constitute a continuing threat to society" (Texas Code of Criminal Procedure, 2013), where "society" includes the prison society (i.e., guards and other inmates; Texas Defender Service, 2004; see also *Estrada v. State*, 2010). In other words, the State of Texas requires that the jury predict dangerousness in convicted criminals.

The constitutionality of Texas's sentencing procedures was upheld by the United States Supreme Court in *Jurek v. Texas* (1976). Justice John Paul Stevens stated in his opinion,

> It is, of course, not easy to predict future behavior. The fact that such a determination is difficult, however, does not mean that it cannot be made. Indeed, prediction of future criminal conduct is an essential element in many of the decisions rendered throughout our criminal justice system. The decision whether to admit a defendant to bail, for instance, must often turn on a judge's prediction of the defendant's future conduct. And any sentencing authority must predict a convicted person's probable future conduct when it engages in the process of determining what punishment to impose. For those sentenced to prison, these same predictions must be made by parole authorities. The task that a Texas jury must perform in answering the statutory question in issue is thus basically no different from the task performed countless times each day throughout the American system of criminal justice. (pp. 274–276)

This is a daunting task assigned to twelve laypersons, plus research has shown that capital murder defendants may not be anymore a threat to prison society as any other prisoner (e.g., Sorensen & Pilgrim, 1999; Reidy, Cunningham, & Sorensen, 2001).

Typically, the jurors will have the help of an "expert" witness. In 1981, The United States Court of Appeals for the Fifth Circuit held,

> Texas may not use evidence based on a psychiatric examination of the defendant unless the defendant was warned, before the examination, that he had a right to

> remain silent; was allowed to terminate the examination when he wished; and was assisted by counsel in deciding whether to submit to the examination. (*Estelle v. Smith*, 1981, p. 461)

In 1983, the United States Supreme Court ruled that psychiatrists are competent to give their opinion on whether a defendant will commit crimes in the future or is a danger to the community: "The suggestion that no psychiatrist's testimony may be presented with respect to a defendant's future dangerousness is somewhat like asking us to disinvent the wheel" (*Barefoot v. Estelle*, 1983, p. 896). The Court cited the *Jurek* ruling in its defense of clinical predictions of dangerousness, saying,

> [I]f it is not impossible for even a layperson sensibly to arrive at that conclusion, it makes little sense, if any, to submit that psychiatrists, out of the entire universe of persons who might have an opinion on the issue, would know so little about the subject that they should not be permitted to testify. (pp. 896–897)

But as Faigman et al. (2013) point out, the ruling assumes the *Jurek* decision was correct in determining laypersons' ability to predict violence and that experts are on the "same constitutional plain" (p. 122). Furthermore, the *Barefoot* ruling was despite the American Psychiatric Association's amicus curiae brief:

> Psychiatrists should not be permitted to offer a prediction concerning the long-term future dangerousness of a defendant in a capital case, at least in those circumstances where the psychiatrist purports to be testifying as a medical expert possessing predictive expertise in this area. Although psychiatric assessments may permit short-term predictions of violent or assaultive behavior, medical knowledge has simply not advanced to the point where long-term predictions—the type of testimony at issue in this case—may be made with even reasonable accuracy. The large body of research in this area indicates that, even under the best of conditions, psychiatric predictions of long-term future dangerousness are wrong in at least two out of every three cases. (p. 3)

Dershowitz (1967) discusses the "the social costs incurred by the increasing involvement of the psychiatrist in the administration of justice" (p. 370). As he recounts,

> My own conversations with psychiatrists reveal wide differences in opinion over what sorts of harms justify incarceration. As one would expect, some psychiatrists are political conservatives while others are liberals; some place a greater premium on safety, others on liberty. Their opinions about which harms do, and which do not, justify confinement probably cover the range of opinions one would expect to encounter in any educated segment of the public. But they are opinions about matters which each of us is as qualified to make as they are. (p. 374)

One of the most infamous psychiatrists involved in predicting future dangerousness was Dr. James Grigson, otherwise known as "Dr. Death." Dr. Grigson, an "expert" witness for the prosecution in both *Smith* and *Barefoot*, was a forensic psychiatrist practicing in Texas and was notorious for his willingness to take the stand for the prosecution in capital punishment trials and predicting whether the recently convicted defendant was dangerous. Dr. Grigson claimed he could "predict with 100 per cent certainty that the individuals would engage in future violent acts" (Biel, 1995). Dr. Grigson testified in over 150 capital murder trials, typically testifying for the side of the prosecution (i.e., supporting the death penalty). In 1995, Dr. Grigson was expelled from the American Psychiatric Association and the Texas Society of Psychiatric Physicians for "alleged ethics violations" (Biel, 1995). Often, Dr. Grigson's opinions were given without interviewing the defendant; instead they came in the form of responses to hypothetical scenarios. The documentary film *The Thin Blue Line* (Lipson & Morris, 1988) is about a man named Randall Dale Adams who was wrongly convicted of the murder of a Texas police officer. After the release of the film, which included a confession from David Harris, Adams was released after spending twelve years in prison. Dr. James Grigson served as an expert witness on Adams' trial and predicted Adams would kill again (Rosenbaum, 1990). Despite being exonerated a year earlier, Rosenbaum reported that

> [Grigson] told me that, despite everything, he *still* [original emphasis] has no doubt about it: the ... confession by [David Harris] was a sham ... [Grigson later] testified not only that he believed Randall Dale Adams did the killing, but that he was certain Adams "will kill again." (p. 166)

Adams lived a quiet life until his death in 2010 (Martin, 2011).

Dr. Richard Coons is another notable Texas psychiatrist who testified in more than 50 capital murder trials (Kreytak, 2010). Like Dr. Grigson, Dr. Coons's paid testimonies provided the courts with "expert" opinion on the matter of future dangerousness for recently convicted murderers facing the death penalty. Also like Dr. Grigson, Dr. Coons sided with the prosecution more often than not. In 2010, the Texas Court of Criminal Appeals ruled that Coons's predictions of future dangerousness were not reliable (Kreytak, 2010). In *Coble v. State* (2010), Dr. Coons "forthrightly stated that 'he does it his way' with his own methodology and has never gone back to see whether his prior predictions of future dangerousness have, in fact, been accurate" (p. 279). In *Espada v. State* (2008) he testified,

> [H]e did not know his rate of error; ... his opinion regarding a defendant's future dangerousness was ultimately based on his professional training and experience; ... his methodology was not based on any specific scientific study; ... it is impossible to conduct accurate scientific research regarding capital defendants' future dangerousness because such defendants "go to death row"; ... [and] it is impossible to "get the same level of hard data reliability [about future dangerousness] that you can [get] in [the] hard sciences."

Regarding the admissibility of Dr. Coons's testimony, the Texas Court of Criminal Appeals said,

> [W]e discern no abuse of discretion on the part of the trial court in its admission of Coons's testimony. Given the arguments, information, and evidence before the trial court at the time it ruled, the trial court could have reasonably concluded that psychiatry was a legitimate field of expertise, that predicting future dangerousness

> was within the scope of psychiatry, and that Coons's testimony would properly rely upon the principles involved in psychiatry. Coons testified that he was an experienced psychiatrist, that psychiatrists are called upon to make predictions of future dangerousness "all the time," and that they do so utilizing such factors as he set forth. ... The fact that Coons did not know his rate of error is not dispositive. (*Espada v. State*, 2008)

Reid (2001) summarizes his point well in commenting on psychiatrists providing expert testimony in capital punishment cases:

> There are lots of appropriate places in which to voice one's opinions about social, moral, or political aspects of the death penalty, but if one has been asked to give psychiatric expert testimony, concerning psychiatric characteristics of a defendant or other people, the courtroom is not one of them. (p. 216–217)

2.2.2 "Operation Baxstrom"

In 1961, a New York convict named Johnnie Baxstrom, while in prison, was certified as insane by a prison physician and shortly thereafter transferred to Dannemora State Hospital for criminals with mental illnesses. After serving out his prison sentence, a petition requesting that he be civilly committed was granted. Baxstrom was denied transfer to a civil mental hospital and was kept at Dannemora without a jury review, denying him equal protection of the law. In 1966, the United States Supreme Court ruled that Mr. Baxstrom was "entitled to a hearing under the procedure granted all others by ... the New York Mental Hygiene Law to determine whether he is so dangerously mentally ill that he must remain in a hospital maintained by the Department of Correction" (*Baxstrom v. Herold*, 1966, p. 383). (For a thorough description of the life of Johnnie Baxstrom and the trial proceedings, see Steadman & Cocozza, 1974.)

Over a six month period following the Supreme Court ruling in *Baxstrom v. Herold*

(1966), "Operation Baxstrom," as it was called,[2] involved the transfer of slightly less than 1000 patients from the Department of Correction to Department of Mental Hygiene institutions (Hunt & Wiley, 1968). Cocozza and Steadman (1975; see also Steadman & Cocozza, 1974) followed 967 of the patients, all deemed dangerous, and observed that only 26 of them "exhibited sufficiently violent behavior" that "justif[ied] their return to hospitals for the criminally insane" (Cocozza & Steadman, 1975, p. 1093). Of the 98 patients that were released into the community, only two reported cases were offenses the authors considered to be dangerous (Cocozza & Steadman, 1975). The results of their study led the authors to conclude the "inability of psychiatrists to accurately predict dangerousness" (Cocozza & Steadman, 1975, p. 1093).

Using clinical diagnoses, Kozol, Boucher, and Garofalo (1972) determined that "dangerousness [could] be reliably diagnosed" (p. 392). The authors, in summarizing dangerous clinical diagnoses state,

> The diagnosis of dangerousness is based on inquiry and examinations that extensively pursue areas of concern not fully dealt with in routine psychiatric assessment. There are no rigid criteria of dangerousness; there are only clues gleaned from a meticulous inquiry into multiple aspects of the personality. We have developed these clues out of painstaking years of trial and error, in the course of which we have developed frames of reference for investigation of the personality. Out of these investigations emerges our clinical *prediction* [original emphasis] as to the patient's future behavior. (p. 383)

Monahan (1973) noted that 65% of the patients Kozol et al. (1972) predicted as dangerous were not dangerous:

> When an extraordinarily thorough clinical examination by at least five mental health professionals combined with an extensive social history and psychological test bat-

[2]In a memo on June 24, 1966, from Robert C. Hunt, MD., Assistant Commissioner, N.Y. State Department of Mental Hygiene

> tery is inaccurate in two out of every three predictions of dangerousness, one cannot conclude that 'reliable diagnosis' of dangerousness has been achieved. (p. 419)

Clinical (expert) prediction, which is based on memory, "gut" feelings, intuition, personal judgment, and the like, should not be used for assessing and predicting future dangerousness when actuarial methods are available; there is more than enough evidence to support the notion that clinical prediction is inferior to or, at best, no worse than statistical prediction (e.g., Sarbin, 1943; Meehl, 1954; Dawes & Corrigan, 1974; Dawes, 1979; Dawes, Faust, & Meehl, 1989; Grove, Zald, Lebow, Snitz, & Nelson, 2000; Ægisdóttir et al., 2006; with respect to predicting criminal behavior, see R. E. Thompson, 1952; Glaser, 1955; J. S. Carroll, Wiener, Coates, Galegher, & Alibrio, 1982; Holland, Holt, Levi, & Beckett, 1983; Mossman, 1994a; Gardner, Lidz, Mulvey, & Shaw, 1996a; G. T. Harris, Rice, & Cormier, 2002; McMillan, Hastings, & Coldwell, 2004; Hilton, Harris, & Rice, 2006). As Dawes (1986) says, "The fact that two dimensions are correlated in nature ... does not imply that they are not psychologically independent and distinct for the perceiver or judge" (p. 14).

2.3 Statistical Prediction

Statistical (actuarial, mechanical) prediction eliminates the human judgment component; predictions are based on empirically-informed relationships. Although some early papers focused on the relationship of intelligence and juvenile delinquency (Pintner & Reamer, 1918; Clark, 1920), statistical prediction in the legal/penal system really began by attempting to predict parole violations (Warner, 1923). In critiquing Warner's research, H. Hart (1923) put the author's results through rigorous analyses (statistical significance testing) while simultaneous laying the statistical framework for which parole prediction should be analyzed.

2.3.1 Burgess Method

Burgess (1928) was one of the first to suggest that empirical evidence should be constructed and used in parole decision making. Burgess proposed 12 significant factors to be used in predicting parole violations: type of offense, parental and marital status, criminal type, social type, community factors, statement of trial judge and prosecuting attorney, previous criminal record, work record previous to commitment, punishment record in institution, intelligence rating, psychiatric personality type, and psychiatric prognosis. He also noted the parole violation rate associated with the factors and examined how the factors "might be combined so as to give more certainty of predictability than any factor taken separately" (p. 247). This led to 21 factors on which a parole-eligible man would be graded and compared to the average case, in turn leading to probabilities of failing parole; the higher the score, the less likely to violate parole. These were charted in an "expectancy" table; as Burgess states, "The prediction would not be absolute in any given case, but, according to the law of averages, would apply to any considerable number of cases" (p. 246). He also warns, "Although statistical prediction is feasible on the basis of data now accessible, exclusive reliance should not be placed on this method" (p. 249). Tibbitts (1931), similar to Burgess (1928), examined what factors could be used in predicting parole violations, including several more factors (23 total) and three times the sample size ($N = 3,000$). Tibbitts (1931) concludes that methods for predicting parole violation "should enable the work of parole administration to be placed on a scientific basis" (p. 49).

In his article *Prediction of Criminality*, Hakeem (1945) provides an extensive bibliography of prediction of criminality, a list that includes over 60 journal articles, books, unpublished manuscripts, and technical reports. Just 17 years after Burgess' paper, it was clear that "considerable attention ha[d] been given to the problem of the application of prediction techniques to criminological data and more especially to the possibility of predicting success or failure on parole" (Hakeem, 1945, p. 31). By the 1940s, one state, Illinois, was

using actuarial methods in its penal system (Hakeem, 1945).

Although much of the research discusses the numerous factors that may be associated with delinquency or parole violations and encourages the use of predictive measures (e.g., Lanne, 1935; Monachesi, 1939), others were skeptical. From the beginning, criticism has followed the use of actuarial methods in predicting future behavior. S. A. Rice (1928) questions the assumed static variables, focusing on the factors used by Burgess (1928), and also questions Burgess's factor selection method. In his paper, Huff (1936) provides an in-depth examination of previous parole prediction studies "to determine their sufficiency as revealed by [the scientific method]" (p. 207). The first issue posed by Huff was the differing opinions regarding the time period for parole violations; in noting the inconsistent views he asks, "How can we secure comparable results if we do not have a common measure of success?" (p. 208). Huff also questions the reliability of the data gathered (i.e., the factors associated with parole violations) as well as their validity. He next questions the generalizability of the associated factors: "Are the norms suitable in terms of a group which is unrepresentative of the general population at the time of classification, and which is a member of the general population during the time of testing?" (p. 211). His third issue, similar to an issue posed by S. A. Rice (1928), regards the static measures used: "It should be inconceivable that a known unfavorable fact would be allowed to remain uncorrected or at least unmodified" (p. 211). His fourth and final—and arguably most important—question focuses on the replicability of results: "We think the premise that [previous results] will be repeated is not sufficiently established by the data offered" (p. 211). Huff concludes,

> Before we can believe parole prediction is a science, we will need to see existing studies applied to new and controlled situations. Adequate data must be much more inclusive than at present. ... Efforts to do so are praiseworthy, but, to date, are not convincing. (pp. 212–213)

These four questions, posed in 1936, are still relevant in the present research regarding prediction of future behavior.

2.3.2 The Glueck Method

Two influential figures in the use of actuarial methods during the mid-20th century were Sheldon and Eleanor Glueck. During their long careers at Harvard Law School, this husband and wife pair wrote numerous books on crime and delinquency based on four longitudinal studies; one of the more notable was entitled *Unraveling Juvenile Delinquency* (1950). Their study included a group of 500 delinquent and 500 non-delinquent white boys aged 9–17; each boy from the delinquent group was matched to a boy in the non-delinquent group with respect to age, general intelligence, national (ethnico-racial) background, and residence in (Boston-area's) underprivileged neighborhoods (Glueck & Glueck, 1950). As they defined it,

> [D]elinquency refers to repeated acts of a kind which when committed by persons beyond statutory juvenile court age of sixteen are punishable as crimes ... except for a few instances of persistent stubbornness, truancy, running away, associating with immoral persons, and the like. (p. 13)

Glueck and Glueck (1950) chose five factors of social background (herein referred to as the Glueck Social Prediction Scale [SPS]) that "differentiate between boys who are potential delinquents and those who are not" (p. 259): discipline of boy by father, supervision of boy by mother, affection of father for boy, affection of mother for boy, and cohesiveness of family. Each of these five factors had subcategories and associated "failure" scores (e.g., overstrict or erratic discipline of boy by father was given a weighted failure score of 72.5). The failure scores are to be summed across the five factors; the range of possible scores is 116.7–414. Glueck and Glueck (1950) developed predictive tables from these failure scores, reducing the blocks of scores from 7 to 4, 3, and 2 allowing the reader to use his or her choice of the score ranges to determine a probability of delinquency. For the two-block prediction table, a score of 250 represents the cutscore or discriminative point for prediction: Those with a score of 250 or greater represent a greater likelihood of delinquency than those scoring below

250. Glueck and Glueck (1968) developed two other prediction tables; one was based on five character traits of the Rorschach test, the other on five personality traits derived in a psychiatric interview. In addition, they combined the data gathered from the three tables to create bivariate comparisons (based on low and high chance of potential delinquency) of the three tables. In terms of practicality, the five social factors held a large advantage. The authors emphasized their prediction tables "should not be used mechanically nor substitute judgment ... [t]hey are designed to aid the clinician in the always difficult task ... of individualization" (p. 269).

The Gluecks followed up their study when the boys approached age 25, and again when they approached 31 (Glueck & Glueck, 1968). Of the 1000 boys, the Gluecks were able to collect data on 880 of them (438 delinquents and 442 nondelinquents); their followup book, *Delinquents and Nondelinquents in Perspective*, details these 880 cases (Glueck & Glueck, 1968). The main goal of their research was to determine if delinquency in childhood led to criminality in adulthood. Glueck and Glueck (1968) determined that most nondelinquent boys continued to live their lives as law-abiding citizens and, although many of the delinquent boys committed crimes in their adult life, the percentage who committed crimes before the first follow-up time was far more than the percentage before the second follow up (77.4% versus 51.8%). The authors conclude this reduction over time "suggests the important role of *delayed maturation* [original emphasis] in the abandonment of major crime" (Glueck & Glueck, 1968, p. 151). The notion that age and recidivism are closely related is still extremely relevant in today's research (e.g., Hanson, 2002; Lussier, Tzoumakis, Cale, & Amirault, 2010; Helmus, Thornton, Hanson, & Babchishin, 2012); an age-related variable is almost always found in actuarial models for recidivism and dangerousness.

Following the publication of *Unraveling Juvenile Delinquency*, several attempts were made to validate the Glueck SPS. R. E. Thompson (1952) used the Glueck SPS to predict juvenile delinquency and, in addition, to compare these predictions to that of three experts (one psychiatrist and two social workers). The data were obtained from the Cambridge-

Somerville Youth Study (CSYS; see Powers & Witmer, 1951). The CSYS included 782 boys all rated by the three experts on an 11-point scale, ranging from −5 to +5 where a negative score represents a boy more likely to become delinquent (labeled as "difficult"); a positive score reflects not becoming a delinquent (labeled "average"). Of the original 782 boys, 650 were matched based on several variables forming 325 pairs; within each pair, one boy was assigned to a treatment group, the other to the control group. Thus, some of the average boys were assigned to the treatment group and some of the difficult boys were not (for further details, see Powers & Witmer, 1951). R. E. Thompson (1952) "randomly" (see p. 457) selected 100 cases to be rated by Eleanor Glueck and used in the validation study. The Glueck SPS showed to be far superior than any of the three judges, correctly classifying 91 of the 100 boys (48 of 52 of those in the control group; 43 of 48 in the treatment group). A prediction was made from all four sources (Glueck SPS and the three experts) for 77 of the cases; the best expert prediction was correct in 67.5% of the cases compared to the 89.6% accuracy of the Glueck SPS. In a second validation study, R. E. Thompson (1957) used the Glueck SPS for fifty delinquent boys and fifty delinquent girls, and correctly classified forty-six of the fifty boys and all fifty girls. In the previous study, 20% of the sample was delinquent; here, 100% was.

Black and Glick (1952), around the same time as R. E. Thompson's article, published a monograph entitled *Recidivism at the Hawthorne-Cedar Knolls School*. The Hawthorne-Cedar Knolls (HCK) school is located in Hawthorne, New York; in 1952, it served as a treatment and rehabilitation center for troubled Jewish boys and girls. From a sample of 100 boys discharged at least five years prior, the authors applied the Glueck SPS. (It should be noted that the authors' main purpose was to examine how well the treatment at HCK performed; they used the Gluecks' Recidivism Prediction Scale [RPS; Glueck & Glueck, 1934], and predicted that all 100 boys had at least a 50% chance of recidivating, an indication that the treatment was a success.) Of the 100 boys, 91 would have been predicted to become a delinquents. Similar to R. E. Thompson (1957), the rate of delinquency in the

sample was 100%.

Two items are worth noting for all three validation studies. First, in all the studies the authors received some sort of consultation from the Gluecks. Second, all "predictions" were, in reality, "postdictions"; that is, the classifications of the boys as likely delinquent were made after the fact. In addition to the three studies discussed, the Gluecks *Unraveling Juvenile Delinquency* data were reanalyzed, and mostly confirmed, almost forty years later (Laub & Sampson, 1988).

The Gluecks were not without their critics, especially among the sociologists (Laub & Sampson, 1991). Their biggest opponent was Edwin Sutherland, "widely acclaimed as the dominant criminologist of the 20th century" (Laub & Sampson, 1991, p. 1402). In a Sutherland review of their book *Later Criminal Careers*, he questions their data and methods, stating that "[t]here was no psychiatric examination of the offenders, with a few exceptions, during the second five-year period. The authors do not explain specifically how they secured this second psychiatric classification" (Sutherland, 1937, p. 186). The validity of some of their measures were also called into question (Hirschi & Selvin, 1967). One common criticism of the Gluecks is their atheoretical approach to criminology (e.g., Reis, 1951; Laub & Sampson, 1991). A more scathing critique of the Gluecks' *Unraveling Juvenile Delinquency* was Reis (1951), noting their and others' failure to validate their predictive tables (this paper was published prior to R. E. Thompson [1952] and Black and Glick [1952]): "empirical research and critical evaluation of the Gluecks' work have not shown their tables to yield valid prediction" (p. 118). Reis also notes the rate of delinquency in the Gluecks' sample is exactly 50%: "Unless this is the actual rate in a similar population for which predictions are made, the tables will yield very poor prediction" (p. 118). Reis considers hypothetical results for the Glueck SPS applied to a sample with a more reasonable estimate of delinquency, 10%, and the results are far less convincing.

Shortly after the Gluecks' first published book, *500 Criminal Careers* (Glueck & Glueck, 1930), two studies (Vold, 1931; Monachesi, 1932) compared the two prominent pre-

diction methods at the time: the Glueck's and Burgess's. Vold (1931) found the two methods comparable, but notes Gluecks' method is more tedious than Burgess's. In another study, Monachesi (1945) used Burgess's method to predict outcome (failure/success) of juvenile probation cases with reasonable success. Ohlin and Lawrence (1952) compare Burgess's method with the method proposed by Laune (1936) based on a questionnaire measuring inmates' attitudes, leading to mixed results.

Predictions of juvenile delinquency and parole violations were becoming more prevalent around the middle of the 20th Century. Despite the Gluecks' contributions, instruments for predicting future criminological behavior were dominated by those in the sociological and criminological fields. Previous research supported the notion that relationships between personality and criminality exist (e.g., using the MMPI, see Capwell, 1945; Monachesi, 1948; Schuessler & Cressey, 1950; Monachesi, 1950a; Monachesi, 1950b; however, see Mack, 1969 for contrary evidence and Schuessler & Cressey, 1950, who question the conclusions of such studies relating personality and criminal behavior). Weinberg (1954) notes that despite the research, many of the early prediction instruments lacked personality and other related factors relevant to theories of criminal behavior:

> The findings of prediction are designed usually for administrative purposes and are within the scope of applied theory. Theories of causation usually consider arrest or detection as superfluous in understanding criminal behavior. The findings of prediction—with the exception of work on potential delinquency—operate within the policy framework of probation or parole. Hence detection or arrest may be crucial. Also the predictive instrument does not always test personality dynamics and social dynamics but becomes a function of empirical trial and error outcome for specified items in terms of a given criterion, whether it be arrest or violation of parole or other action. Frequently, these items are taken from records in the files and have been recorded for purposes other than prediction. Hence these items become direct or indirect indexes at best of the pertinent behavioral dynamics of

> criminality. On the other hand, there has been an increasing recognition of the need to use items derived from a consistent theory and then applied to predictive tests. (p. 422)

Hathaway and Monachesi (1951) further note that "[t]he majority of published studies undertaken to discover the personality characteristics of young people who later develop behavioral disorders have been based upon data collected after the individuals became deviant" (p. 469). Hathaway and Monachesi (1951; see also Hathaway & Monachesi, 1952) used the Minnesota Multiphasic Personality Inventory (MMPI) to measure personality on 4,046 ninth graders (approximately aged 14; 1,994 boys, 2,052 girls) in the Minneapolis area (their research was prior to the MMPI-A, the adolescent version of the MMPI, and the authors acknowledged the limitations of interpreting juvenile results on an adult inventory). Three years later, the authors determined, through the local correctional system, that 597 of the 4,046 adolescents (442 boys and 155 girls) of the boys and girls had come into contact with the court or police. They then compared the MMPI profiles of the delinquents and nondelinquents; the authors conclude that "several scales of the MMPI differentiated significantly between delinquents and non-delinquents" (Hathaway & Monachesi, 1952, p. 708).

Ohlin and Duncan (1949) looked at over a dozen studies with readily-constructed data, focusing on a measure they called the *percentage in the error of prediction* (see also Horst, 1941, pp. 258–259, under the section, *The Principle of Maximum Probability*). The percentage in the error of prediction is the difference in error rates of the prediction measure from prediction using total rates, or base rates, divided by the error using total rates. Using their example, suppose that 40.1% of parolees violate the conditions of their parole. Then prediction of a random individual suggests stating every parolee will be a success so that the method is correct 59.9% of the time. Suppose a prediction method has an error percentage of 32.5%; the difference between the method and total rates prediction is $(40.1 - 32.5 =)$ 7.6% and the percentage in the error of prediction is $^{7.6}/_{40.1} = 19.0\%$. The authors state,

> In only two instances do the reductions in error [i.e., the percentage in the error of

> prediction] greatly exceed 25 per cent, and, of these two, one is based on a small number of cases. ... [I]t is quite clear that routine application of these techniques to the types of data usually secured is in no sense a guaranty of substantial improvements in prediction over the crudest method available—prediction from total rates. (p. 445)

The research throughout the 1950's and 60's continued to support prediction methods in penal systems and these methods became more sophisticated. Duncan, Ohlin, Reiss, and Stanton (1953) discuss six rules previously suggested in the literature for optimizing cutscores based on cost and utility, where cost is associated with false negatives and utility with true negatives. Glaser (1955) uses one of Duncan et al.'s measures to compare actuarial predictions to predictions made by criminologists and psychiatrists for parole success. His results suggest that criminologists' predictions are superior to those made by psychiatrists, but actuarial methods outperform both types of clinical prediction. Kirby (1954) is the first to develop a predictive model for parole violations using discriminant function analysis, improving upon the methods of Burgess and the Gluecks. Mannheim and Wilkins (1955) use multiple linear regression to predict criminal recidivism; it is the first such study conducted in England. However, not all results were positive; for instance, in an application based on results from previous literature, Schlesinger (1978) used previously-reported significant predictors to predict dangerousness in juveniles but the results failed to support accurate predictions of dangerousness.

With advances in prediction models, Rose (1966) questioned what one does—whether it be an institution or a legal system—with a prediction for a specific individual. Rose (1966) states,

> [T]he prediction does not indicate ... which form of treatment action is to be more successful, and thus what [one] should actually do. ... These are the kind of questions which must be faced, and unless they are, the application of prediction scores to individuals is not likely to be of much use. At the present moment our

> effectiveness at describing and identifying different treatment techniques is very low, and we cannot, therefore, produce the kind of significant relationships which are the basis of a prediction table. It may be that the criterion of recidivism also is too gross for this type of operation, which should be much more based upon some measure of personality change—something else which we are bad at measuring. (p. 30)

2.4 Predicting Dangerous Behavior

"It has been suggested that dangerousness, like beauty, lies in the eye of the beholder" (Shah, 1978, p. 224). Predicting dangerousness begins with defining dangerousness; in his influential article, Shah (1978) defines it as follows:

> a propensity (i.e., an increased likelihood when compared with others) to engage in dangerous behavior. *Dangerous behavior* refers to acts that are characterized by the applications of or the overt threat of force and that are likely to result in injury to other persons. (p. 224)

Shaw (1973) states,

> The problem of 'dangerousness' is its definition. There seems to be three terms used synonymously: 'dangerousness', 'aggression', and 'violence'. ... 'violence' and 'aggression' denote action, while 'danger' denotes a relationship. (p. 269)

In *State v. Krol* (1975), the United States Supreme Court weighed in on the matter: "Dangerous conduct is not identical with criminal conduct. Dangerous conduct involves not merely violation of social norms enforced by criminal sanctions, but significant physical or psychological injury to persons or substantial destruction of property" (p. 259). In *Overholser v. Russell* (1960), the opinion of the Court stated,

> We think the danger to the public need not be possible physical violence or a crime of violence. It is enough if there is competent evidence that he may commit

> any criminal act, for any such act will injure others and will expose the person to arrest, trial and conviction. There is always the additional possible danger—not to be discounted even if remote—that a non-violent criminal act may expose the perpetrator to violent retaliatory acts by the victim of the crime. (p. 198)

2.4.1 *Negro v. Dickens* (1965)

In 1944, a man named Joseph Negro was committed—pretrial—to Matteawan State Hospital in New York; Negro plead not guilty for second-degree robbery, first-degree grand larceny, second-degree assault, and felonious prison escape. He was declared incompetent to stand trial and ordered to remain at the institution until "he [was] no longer in such state of idiocy, imbecility or insanity as to be incapable of understanding the charge against him or of making his defense thereto" (*Negro v. Dickens*, 1965, p. 408). Negro spent 21 years at Matteawan (an amount of time that exceeded the maximum sentence had he stood trial and been convicted); the District Attorney of New York would have dismissed of the indictment if the superintendent of Matteawan recommended the patient could be transfered to a civil hospital. The superintendent would not recommend dismissal until the Department of Mental Hygiene determined the appropriateness of transfer to a civil hospital. The Department of Mental Hygiene, presiding over the New York civil hospitals, believed that Matteawan, being under the Department of Correction's jurisdiction, was the one to make the decision. Upon hearing this, the superintendent of Matteawan, in a letter to Negro's counsel, stated:

> I would resent any member of our psychiatric staff at this hospital, making the decision that any patient is suitable for care in a civil hospital. This is a determination that should be made and must be made by the Department of Mental Hygiene. Our staff is working in a closed hospital and they cannot be the authority for the open civil hospitals of New York State. ... [R]efer this matter to the Department of Mental Hygiene for a decision. (p. 410)

Neither side was willing to take a stance. The Court stated,

> We are therefore confronted, upon this record, with a situation in which two official agencies function within medical areas and each during a protracted exchange disavows the responsibility of furnishing the District Attorney with a statement as to whether petitioner's transfer to a civil hospital would be in the public interest. In his affidavit upon the motion the District Attorney equated the issue of "danger to the community" with "the public interest" as a controlling factor as to whether the indictment should be dismissed. Obviously, the District Attorney cannot decide this issue without authoritative advice from men of medicine. (p. 411)

As Morris (1967) contends,

> The ... case of *Negro v. Dickens* aptly demonstrates the inability of Department of Correction psychiatrists to formulate an expert opinion as to whether a patient in a Department of Correction mental institution, who is not dangerous to that institution, would be a danger if transferred to a Department of Mental Hygiene hospital. (pp. 677–678).

Until the late 1960's, criminal prediction primarily focused on juvenile delinquency and parole violations. But expert predictions of dangerousness and violence, especially following Operation Baxstrom and the subsequent research (e.g., Cocozza & Steadman, 1975; Steadman & Cocozza, 1974), were losing credibility whereas predictions by actuarial methods started to garner interest; however, few actuarial methods existed, and Wolfgang (1969) criticizes the inability of these methods to distinguish violent and nonviolent prisoners. He calls for improvements in the classification of violent criminals and argues this was necessary so that appropriate treatments could be applied (e.g., sending an "essentially nonviolent" criminal convicted of a violent offense to a minimum- or medium-security prison rather than a maximum-security prison).

2.4.2 Preventive Detention

Shortly after taking office, President Richard Nixon gave a statement regarding violence and crime in the country, and particularly in the nation's capital. Nixon stated, "in the midst of a crime crisis, immediate steps are needed to increase the effectiveness of the police and to make justice swifter and more certain" (Nixon, 1969). Nixon proposed action in 12 major areas to combat the so-called crisis; one such area was bail reform. Nixon stated,

> Increasing numbers of crimes are being committed by persons already indicted for earlier crimes, but free on pretrial release. This requires that a new provision be made in the law, whereby dangerous hardcore recidivists could be held in temporary pretrial detention when they have been charged with crimes and when their continued pretrial release presents a clear danger to the community.

What constituted a "clear danger to the community" was not discussed by Nixon. Although predictive techniques were becoming more popular and widespread, others were questioning the ethical implications behind their use. As Wenk, Robison, and Smith (1972) note,

> Concern about violence will inevitably lead to the development of special treatment programs, but the majority of persons placed in such programs must be false positives—persons who would not commit the act which the program is designed to prevent. ... Those who argue that treatment cannot harm the person who does not need it and those who would warp the definition of "need" are obviously ignorant of the effects of social stigma and of the difficulty of administering corrective interventions without social stigma as a result. (p. 402)

Many were opposed to the general idea of *preventive detention* (i.e., commitment of individuals based on suspicion to commit a crime) that Nixon supported. Preventive detention, or what Dershowitz (1970) refers to as "the prediction-prevention strategy," requires an assessment of an individual and, thus, is based on some sort of prediction of whether

the individual is likely to skip bail, commit a crime, pose a clear danger to the community, and so forth. Dershowitz (1970) notes that the "two prediction-prevention devices employed most widely in the United States today are: 1) denial of pre-trial release to persons charged with, but not yet convicted of, crimes; and 2) involuntary hospitalization of the mentally ill" (p. 29). Some contend that preventive detention is unconstitutional and specifically violates the Fifth and Eighth Amendments: "nor [shall a person] be deprived of life, liberty, or property, without due process of law" (U.S. Const. amend. V) and "Excessive bail shall not be required, nor excessive fines imposed, nor cruel and unusual punishments inflicted" (U.S. Const. amend. VIII).

As Foote (1970b) argues,

> [B]oth money bail, which in practice has amounted to preventive detention for the poor, and preventive detention of the dangerous as explicitly authorized in the District of Columbia, should be held unconstitutional under the due process and equal protection clauses. The same result should also follow under the Eighth Amendment forbidding excessive bail, where the court will have to resolve the ambiguity of a carelessly drafted clause. (p. 4)

Foote continues,

> There is no plausible theory on the drawing boards of social science, let alone any empirical evidence, which supports the simple-minded assumption that specific criminality can be predicted in advance without error so great that its application would be arbitrary in the constitutional sense. The required prediction cannot be achieved by any existing clinical or scientific expertise, let alone by scientifically untrained prosecutors and judges in the hurlyburly of our municipal courts. (p. 4)

In determining dangerousness analytically, Foote states there are distinct questions that need to be answered:

> (1) What kinds of anticipated criminal conduct are harmful enough to justify the imposition of pretrial detention?; ... (2) What are the probabilities that pending

> trial the defendant would engage in each kind of specified conduct?; (3) How is one to combine the weight assigned to the harm which the defendant might do ... with the probability that he would do it ... in order to rate him on a scale from most serious risk to least serious?; ... [and] (4) How high a risk is high enough to justify jailing someone between arrest and trial despite the empirical proof that detention adversely influences the outcome of the case? (p. 8)

Foote emphasizes this last point as the most crucial and notes the trade-off between degrees of risks: detaining few defendants with a decreased number of crimes prevented (high risk) or detaining many defendants with an increased number of defendants detained who would not have committed a crime (low risk). Foote (1970a) argues that "[j]udges and psychiatrists who support preventive detention assume that a mistaken identification of one actually safe person who is predicted to be dangerous is much less serious than the release of one actually dangerous person" (p. 53).

Another common argument among the opponents of preventive detention is concerned with the number of false positives (i.e., detaining a nondangerous individual incorrectly predicted to be dangerous). As Dershowitz (1971) claims,

> [T]here is no evidence that there is any device or human being that can predict short-term violence with 50–60 per cent accuracy. I cannot personally conceive, at least at the present time, of any predictive criteria that would correctly spot any significant number of violent crimes without requiring the confinement of at least twice—probably far more—but at least twice as many false positives as true positives. Accordingly, under any realistic preventive detention criteria, it will always be more likely than not that a given detainee has been erroneously confined. (p. 564)

John N. Mitchell, U.S. General Attorney under President Nixon, arguing for the constitutionality of preventive detention says, in defense of the Fifth Amendment, states

> [T]he due process clause of the Constitution does not prohibit pretrial detention in criminal cases. Its requirements are those of reasonableness—the restraints imposed

> on the liberty of an accused must be reasonable when balanced against society's acknowledged interest in preventing commission of further crimes while the defendant is awaiting trial. (Mitchell, 1969, pp. 1234–1235)

In discussing the Eighth Amendment, Mitchell says,

> [S]ince the [E]ighth [A]mendment when adopted clearly permitted pretrial detention for capital crimes because of danger to the community, it should not today prohibit pretrial detention for such dangerous crimes merely because they are no longer capital for reasons completely unrelated to their dangerousnes. (p. 1230)

With respect to predicting dangerousness, Mitchell, rather definitively, states,

> [T]he conclusion is inevitable that statistical evidence which permits predictability with precise mathematical accuracy is not constitutionally necessary to warrant confinement on grounds of dangerousness. Instead, it is sufficient to place reliance, as is the practice in the law, on the insight and experience of trial judges applying appropriate qualitative standards. (p. 1241)

In other words, the Honorable John N. Mitchell deemed trial judges capable of predicting future dangerousness.

M. L. Cohen, Groth, and Siegel (1978) argue that clinical prediction of dangerousness "is possible" (p. 34). They cite two articles to support this argument: the first being the aforementioned article by Kozol et al. (1972); the second, an article by Hodges (1971) on the effectiveness of Maryland's indeterminate sentence law. The authors state that the two studies were able to "predict[] not only dangerousness but also nondangerousness" (p. 35). However, this argument was not generally accepted (e.g., Steadman, 1973; Diamond, 1974; Cocozza & Steadman, 1978; for a particularly detrimental review, see Ennis & Litwack, 1974). Specific issues with respect to the Kozol et al. article were previously presented; Alan Stone, in a discussion immediately following the Hodges article, succinctly summarizes his

opinion: "[T]his study presents no evidence that justifies its overblown conclusions or the existence of Maryland's law" (p. 295). But despite evidence against of their ability, predictions of future dangerousness were still commonly made by psychiatrists and psychologists. Even the American Psychiatric Association advised that "[p]sychiatric expertise in the prediction of 'dangerousness' is not established and clinicians should avoid 'conclusory' judgments in this regard" (1983, p. 33). If expert psychiatrists fail to predict dangerousness with acceptable accuracy, how can one expect a judge to fair any better, as John N. Mitchell suggested? Mitchell brushed the issue aside stating, "due process of law requires fundamental fairness, not perfect accuracy" (p. 1242).

Tribe (1970) disagrees with many of Mitchell's (1969) arguments. With respect to the fifth amendment, he says,

> [President Nixon's] proposal says absolutely nothing about the sort of harm the defendant must be found likely to commit before he can be imprisoned pending trial ... [thus] confinement is authorized so long as the accused falls into any detainable category ... [and n]o tenable concept of due process could condone a balance that gives so little weight to the accused's interest in pretrial liberty. (p. 381)

Tribe also suggests that pretrial detention portrays the defendant as guilty before the trial even begins. With regards to Mitchell's defense of the pretrial preventive detention and the Eighth Amendment, Tribe says,

> Attorney General [Mitchell] properly concludes that the excessive bail clause cannot be read to imply an absolute right to pretrial release for all defendants. This conclusion implies only that some governmental interests may justify the denial of pretrial liberty, and leaves open the question of which interests have this character. It is plainly a *non sequitur* [original emphasis] to conclude, as does the Attorney General, that the excessive bail clause leaves Congress entirely free to establish the circumstances under which pretrial release may be withheld. (p. 399)

On July 29, 1970, President Nixon signed the District of Columbia Court Reform and Criminal Procedure Act (1970) saying, "We want to make Washington, D.C., an example of respect for law and of freedom from fear, rather than an example of lawlessness" (Nixon, 1970). Of particular interest are the provisions of pretrial preventive detention outlined in 23 D.C. Code §§ 1322 and 1323, allowing the judicial officer to assess the defendant's dangerousness and authorize pretrial preventive detention of non-capital defendants (Rauh & Silbert, 1970). The constitutionality of the preventive detention provisions of the District of Columbia Court Reform and Criminal Procedure Act was put on trial in *Dash v. Mitchell* (1965; the primary defendant, Mitchell, being the aforementioned Attorney General John N. Mitchell). In addition to citing violations of the fifth and Eighth Amendments, the plaintiffs also argued the act was in violation of the Sixth Amendment which states,

> In all criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the state and district wherein the crime shall have been committed, which district shall have been previously ascertained by law, and to be informed of the nature and cause of the accusation; to be confronted with the witnesses against him; to have compulsory process for obtaining witnesses in his favor, and to have the assistance of counsel for his defense. (U.S. Const. amend. VI)

The case was eventually dismissed on grounds that "all of the plaintiffs in this case, for one reason or another, fail[ed] to present a claim which c[ould] be reached on the merits" (p. 1303).

In addition, civil commitment began including juvenile offenders. For instance, in regards to juvenile delinquents, the New York Family Court Law states,

> The court shall not direct detention unless available alternatives to detention, including conditional release, would not be appropriate, and the court finds that unless the respondent is detained there is a *serious* [emphasis added] risk that he or she

> may before the return date commit an act which if committed by an adult would constitute a crime. (§ 320.5(3)(a)(ii) New York Family Court Law, n.d.)

The provision was brought to the United States Supreme Court in *Schall v. Martin* (1984); it was argued that the provision allows detention of juveniles without due process of the law. The Court upheld the constitutionality of the provision stating, "from a legal point of view there is nothing inherently unattainable about a prediction of future criminal conduct" (p. 278). Ewing (1985) took particular offense to this ruling:

> In *Schall v. Martin*, the Supreme Court held that, in the interest of crime prevention, juveniles merely accused of crimes may be incarcerated before trial, indeed even before a finding of probable cause, simply on the basis of judicial predictions of criminal conduct. Empirical research indicates that such predictions are more likely to prove wrong than right and suggests that this likelihood of error cannot be reduced appreciably by the imposition of procedural safeguards. In light of this research, it seems clear that many if not most juveniles detained on the basis of predictions of criminal behavior will be erroneously identified as potential criminals or recidivists and be needlessly incarcerated. (p. 225)

2.4.3 The Difficulty of Prediction

Monahan (1977) asks, "Can one . . . imagine a Human Subjects Committee approving a study in which one third, or even one tenth, of the subjects may commit assault or murder as a result of the experiment?" (p. 364). He then points out "that the current social policy of confining mentally ill persons on the ground of an admittedly untested prediction of violence is *itself* [original emphasis] an 'experiment'" (pp. 364–365).

Scott (1977) states,

> Prediction of dangerousness is particularly difficult because: dangerousness is the resultant of a number of processes which occasionally may be synergistic amounting to more than the sum of their parts, some within the individual and some in society;

> it is not static; key factors are the individual's adaptiveness, resistance to change, and his intention ... a common mistake is to confuse recidivism with dangerousness, they are not necessarily the same and may be combined in various patterns. (p. 128)

In discussing why clinicians' predictions of dangerous behavior continue to be taken seriously, Berger and Dietrich (1979) posit that "the attitude [is] that any formal procedure, no matter how inadequate, is better than no formal procedure at all" and "that many persons do not clearly understand some fundamental problems in such an endeavour" (p. 36). Scott (1977) argues that "[p]rediction studies should aim not to replace but to complement the clinical approach, and vice versa" (p. 129).

Megargee (1976) notes the difference and difficulty of prediction from postdiction:

> [E]ven if the validity literature showed that our tests could accurately identify people who have been violent in the past, it would not necessarily mean these tests could predict who will behave dangerously in the future. The violent act itself may have created feelings of guilt or relieved pent-up hostility. Also, people who have been identified as having illegally engaged in such acts are inevitably exposed to a variety of judicial and correctional procedures expressly designed to change their personality structure and dynamics. And change they probably do, although the nature of these alterations may be quite different from what was intended. (p. 10)

B. Rubin (1972) opines, "[I]t is unlikely that dangerousness can be predicted in a person who has not acted in a dangerous or violent way" (p. 405). This idea lead Wenk et al. (1972) to conclude that "[t]he prediction equations contain the seed of self-fulfilling prophecy: those who have been noticed before will be noticed again." In summarizing their research, Wenk et al. says, "Our demonstration of the *futility* [original emphasis] of such prediction should have consequences as great for the protection of individual liberty as a demonstration of the utility of violence prediction would have for the protection of society" (p. 402).

Wenk et al. (1972) note the difficulties in predicting violence, even using more sophisticated multivariate techniques, due to the low prevalence of violence and the fact that not all violent acts are reported and the ones that are can vary dramatically. Megargee (1976) also notes the issue of predicting a rare event. Because of the possibly low prevalence of violence, the number of false positives would be high, as Stone (1975) notes:

> [I]f dangerousness is the sole criterion for civil commitment or other preventive detention, and if an empirical study demonstrates violence is a rare event (low base rate), then even if we had a very good predictive technique or device, we would end up confining many more false than true positives.

Shah (1978) argues this idea as well:

> Invariably, such predictions [of low base-rate events] are accompanied by rather huge rates of "false positive" errors; that is, the great majority of the persons predicted as likely to engage in future violent behavior will not display such behavior.

Maden (2003) is more accepting; in his argument for the use of standardized risk assessment methods (i.e., an assessment tool with established norms), he suggests, "Rather than search for the Holy Grail of the perfect risk assessment instrument, there is a strong case for accepting the flaws of an existing scale, which are often outweighed by the benefits of standardisation" (p. 202). Underwood (1979) provides an optimistic viewpoint: "Perhaps no available predictive method is sufficiently accurate to satisfy the high standard of accuracy appropriate for the decision to incarcerate, but it may still be possible to develop one" (p. 1413).

An *illusory correlation* is, as the authors who coined the term define it,

> the report by an observer of a correlation between two classes of events which in reality (a) are not correlated, or (b) are correlated to a lesser extent than reported, or (c) are correlated in the opposite direction than that which is reported. (Chapman & Chapman, 1967, p. 194)

Sweetland (1973) surveyed psychiatrists and naïve participants and found the existence of illusory correlations between personality variables and dangerousness suggesting "that many psychiatrists, and others engaged in making [judgments of dangerousness], are likely to hold beliefs about the relationship between various personality characteristics and the likelihood of dangerous behavior which are of unknown validity" (p. 42). Cunningham and Reidy (1999) present several studies that provide counterintuitive results that lend themselves to illusory correlations.

Kahneman and Tversky (1973) showed that intuitive predictions (e.g., clinical judgment) are influenced by a judgmental heuristic that they called *representativeness* and that this can lead to systematic errors in reasoning. For example, when a psychiatrist or judge believes that a certain group of people are more violent than others, this belief is likely to influence the prediction of future violence and lead to over-prediction for individuals in that group. Diamond (1974) specifically hypothesizes why psychiatrists are likely to over-predict dangerousness:

> If the psychiatrist under-predicts danger, and clears a patient who later commits a violent act, he will be subjected to severe criticism. If, on the other hand, he over-predicts danger, he will suffer no consequence from such faulty prediction, for his prediction might have come true had there been no intervention (such as institutionalization). In general, if the psychiatrist predicts that there is no danger, the feed-back from an erroneous prediction is real and immediate. If he predicts that there is danger, there may be no feed-back, or, if there is, it may not be possible to interpret it in ways which would improve the predictive ability of the psychiatrist. Inevitably, this will result in all concerned doing the "safe" thing: predicting dangerousness, if there are even the most minimal reasons to justify it. (p. 447)

Monahan and Cummings (1974) found that psychiatrists' over-predictions could be partly attributed to the consequence of the prediction to the individual. Hypothetical predictions

of dangerousness were made by two groups of psychiatrists: one group was told that dangerous patients would be hospitalized; the other group believed that the patient would be imprisoned. The psychiatrists were more likely to predict dangerousness if the prediction led to hospitalization rather than imprisonment.

In their assessment of the validity of violence prediction, Monahan and Cummings (1975) conclude, "The empirical fact of the invalidity of violence prediction suggests that neither the interests of society nor the interests of the identified individual are being served by social policies based on the prediction of violence" (p. 162).

2.5 Mental Illness and Violence

Baxstrom v. Herold (1966) brought forth concern over court-ordered treatment of mentally-ill individuals accused—but not necessarily convicted—of a crime, otherwise known as *involuntary commitment*. Involuntary commitment encompasses both criminal and civil commitment; commitment to an institution via a criminal case is *criminal commitment* whereas, in a civil case, it is referred to as *civil commitment*. Civil commitment commonly occurs after an individual serves time for a convicted crime but is still deemed dangerous; it is more common with sexual offenders. The alleged purpose of civil commitment is to protect society from so-called dangerous individuals, and this does not necessarily include treatment.

In the case of *Donaldson v. O'Conner* (1975), the plaintiff, Kenneth Donaldson, was civilly committed in a Florida hospital for nearly 15 years, despite receiving little treatment and not being considered dangerous. In *Donaldson v. O'Conner* (1975), the Court of Appeals distinguished between two types of civil commitment: "police power" (used to commit those who posed a threat of danger toward others) and "parens patriae" (when need for care or treatment is the rationale for confinement). The Court stated, regardless of reason for commitment, "that a person involuntarily civilly committed to a state mental hospital has

a constitutional right to receive such individual treatment as will give him a reasonable opportunity to be cured or to improve his mental condition" (p. 520).

In 1965, Congress passed the District of Columbia Hospitalization of the Mentally Ill Act (HMIA) "[t]o protect the constitutional rights of certain individuals who are mentally ill, to provide for their care, treatment, and hospitalization, and for other purposes" District of Columbia Hospitalization of the Mentally Ill Act (1965). One of the "other purposes" can be found in Sec. 7: Hospitalization Under Court Order. As Cantor and Sherman (1965) put it, "The heart of the Hospitalization of the Mentally Ill Act is contained in its provisions for hospitalization for an indefinite period pursuant to court order" (p. 212). This section of the HMIA allowed "for the judicial hospitalization of any individual in the District of Columbia" with a formal petition:

> [petition needs to] be accompanied (1) by a certificate of a physician stating that he has examined the individual and is of the opinion that such individual is mentally ill, and because of such illness is likely to injure himself or others if allowed to remain at liberty, or (2) by a sworn written statement by the petitioner that (A) the petitioner has good reason to believe that such individual is mentally ill and, because of such illness, is likely to injure himself or others if allowed to remain at liberty, and (B) that such individual has refused to submit to examination by a physician. (District of Columbia Hospitalization of the Mentally Ill Act, 1965)

Similarly, the United Kingdom Parliament passed the Mental Health Act (MHA) of 1983 (Mental Health Act, 1983) that allowed the detention of any patient provided,

> (a) he is suffering from mental disorder of a nature or degree which warrants the detention of the patient in a hospital for assessment (or for assessment followed by medical treatment) for at least a limited period; and (b) he ought to be so detained in the interests of his own health or safety or with a view to the protection of other persons. (Part II, 2.(2))

Szmukler (2003) suggests that the act targets the intersection of two groups (mentally ill persons and potentially dangerous persons) and questions how a law can target a small proportion of those who could be dangerous simply because they are mentally ill (i.e., ignoring those who are dangerous but not mentally ill). He states,

> In the case of mental disorders, we are weighing the benefits to society as a whole against costs that largely fall on a small segment of the population—those with mental illnesses. Given society's long history of prejudice against mentally ill individuals, the threat to this socially-excluded group is very worrying. (p. 206)

Many similar state statutes followed the HMIA, requiring the individual be mentally ill and dangerous; furthermore, an absence of dangerous behavior did not preclude one from commitment. In *Matthew v. Nelson* (1978), the three-judge United States District Court of the Northern District of Illinois concluded,

> [T]here are instances in which a psychiatrist can determine from a psychiatric clinical examination that a mentally ill person is reasonably likely to injure himself or another even though the person's history does not include a recent overt act ... These cases may be relatively few, but they are not so insignificant that they can be discarded in our evaluation. (*Matthew v. Nelson*, 1978, p. 711)

Determining dangerous behavior leading to civil commitment was primarily a responsibility left for the state; in 1976, a court ruling would fault mental health professional for failure to properly report potentially dangerous persons.

2.5.1 *Tarasoff v. Regents of the University of California* (1976)

In September 1967, an Indian man named Prosenjit Poddar came to the United States to attend the University of California at Berkeley as a graduate student; a year later, he met a woman named Tatiana Tarasoff whom he befriended. Poddar developed feelings toward Tarasoff that were not reciprocated, leading Poddar to become emotionally distressed. In

1969, Tarasoff left for South America; Poddar sought psychological help during this time and at one point confided to his psychologist, Dr. Lawrence Moore, that he intended to kill Tarasoff. Dr. Moore recommended that Poddar be civilly committed, feeling that Poddar was a dangerous person. Poddar was detained but released shortly after because he did not appear irrational.

In October of 1969, Tarasoff returned to the United States and on October 27, Poddar fatally stabbed her. Poddar was convicted of second-degree murder but this charge was eventually overturned, conditioned on his returning to India (*People v. Poddar*, 1974). Tarasoff's parents filed a lawsuit against Dr. Moore and the University of California (*Tarasoff v. Regents of the University of California*, 1976). The Court concluded that mental health professionals are responsible for not just their patients, but any individual who may be at risk of harm from their patient. In his opinion, Justice Matthew Tobringer stated, "[T]he public policy favoring protection of the confidential character of patient-psychotherapist communications must yield to the extent to which disclosure is essential to avert danger to others. The protective privilege ends where the public peril begins" (p. 442).

2.5.2 Blackstone's Ratio

In discussing the use of presumptive evidence, Sir William Blackstone famously says, "All presumptive evidence of felony should be admitted cautiously; for the law holds it better that ten guilty persons escape, than that one innocent party suffer" (Blackstone, 1794, p. 713). Many have used the latter part of the quote as an argument against civil commitment and preventive detention; Monahan (1977) disagrees, saying,

> [I]t may be possible ethically to justify short-term commitment even if the predictions of imminent violence on which it is based are less accurate than the long-term research indicates. Paraphrasing Blackstone, it may be better that ten "false positives" suffer commitment for three days than that one "false negative" go free to kill someone during that period. (p. 370)

Slobogin (2009) also disagrees,

> The adage that ten murderers should go free before one innocent person is convicted, although perhaps acceptable as an illustration of our commitment to justice, is much harder to swallow when we know that a sizeable proportion of the ten guilty persons will commit another murder if all of them are let go. (pp. 67–68)

Section 1 of the Fourteenth Amendment declares that no state shall "deprive any person of life, liberty, or property, without due process of law" (U.S. Const. amend. XIV, § 1). The constitutionality of civil commitment was challenged in the case of *Prochaska v. Brinegar* (1979). In the ruling, the Court stated,

> [The appellant Stanley Prochaska] is being restrained of his liberty in that he is not free to come and go at will but such restraint is not in the way of punishment, but for his own protection and welfare as well as for the benefit of society. Such loss of liberty is not such liberty as is within the meaning of the constitutional provision that "no person shall be deprived of life, liberty or property without due process of law." (p. 872)

Laves (1975) argues that civil commitment based on psychiatric predictions of dangerousness was a violation of the Fourteenth Amendment because "psychiatric opinion does not rise to the level of expert testimony" (p. 319).

This notion is not unrecognized; in *Addington v. Texas* (1979), Chief Justice Warren Burger of the United States Supreme Court stated, in his opinion to the Court,

> Whether the individual is mentally ill and dangerous to either himself or others and is in need of confined therapy turns on the *meaning* [original emphasis] of the facts which must be interpreted by expert psychiatrists and psychologists. Given the lack of certainty and the fallibility of psychiatric diagnosis, there is a serious question as to whether a state could ever prove beyond a reasonable doubt that an individual is both mentally ill and likely to be dangerous. (p. 429)

This was also noted in *In re Stephenson* (1977); the Supreme Court of Illinois said,

> [We] believe that proof beyond a reasonable doubt is an inappropriate standard for use in involuntary civil commitment proceedings. Predictions of dangerousness can hardly be beyond a reasonable doubt in the undeveloped framework of the science of psychiatric diagnosis and prediction, for the subjective determinations therein involved are incapable of meeting objective certainty. (pp. 555–556)

In *Addington v. Texas* (1979), the Court argued that "[i]t cannot be said ... that it is much better for a mentally ill person to 'go free' than for a mentally normal person to be committed" (p. 429) and concluded,

> [T]he [proof beyond a] reasonable-doubt standard is inappropriate in civil commitment proceedings because, given the uncertainties of psychiatric diagnosis, it may impose a burden the state cannot meet and thereby erect an unreasonable barrier to needed medical treatment. ... To meet due process demands, the standard has to inform the factfinder that the proof must be greater than the preponderance-of-the-evidence standard applicable to other categories of civil cases. (pp. 432–433)

Thus, the burden of proof required for civil commitment lies somewhere between *preponderance of the evidence* and *beyond a reasonable doubt*, what is referred to as *clear and convincing*. According to Stone (1975),

> The predictive success appropriate to a legal decision can be described in three levels of increasing certainty: preponderance of the evidence, 51 percent successful; clear and convincing proof, 75 percent successful; beyond a reasonable doubt, at least 90 percent successful. (p. 33)

Monahan and Wexler (1978) in their article regarding the standards of proof in civil commitment, state,

> [M]ental health professionals (or actuarial tables) may well be able to prove "dangerousness" beyond a reasonable doubt ... if and only if "dangerousness" is viewed

> as a *probability* [original emphasis] statement, rather than as an *absolute* [original emphasis] claim that violent behavior will occur. (p. 38)

The authors believe that a prediction of dangerousness consists of three distinct assertions: (a) the person has specific characteristics; (b) these characteristics are probabilistically associated with dangerous behavior; and (c) there is a probability of dangerous behavior assigned to that person that warrants civil commitment. They claim it may be possible to prove beyond a reasonable doubt the probabilistic associations of given characteristics with dangerous behavior. The last point, according to Monahan and Wexler (1978), "should be resolved not by 'proof' but by process of legislative policy-making and constitutional 'interest-balancing'" (p. 39).

In 1975, Michael Jones was arraigned on charges of attempted petit larceny, punishable by a maximum sentence of one year in prison in the District of Columbia. The defendant was acquitted by reason of insanity and hospitalized. After several release hearing where he was declared a danger to himself and others, he remained hospitalized for more than a year. In *Jones v. United States* (1983), the United States Supreme Court ruled that "not guilty by reason of insanity is a sufficient foundation for commitment of an insanity acquittee for the purposes of treatment and the protection of society" (p. 366). The Court also dismissed the *clear and convincing* criterion set forth in *Addington v. Texas* for civil-commitment cases and concluded that *preponderance of evidence* was sufficient requirement of proof for the defendant's commitment due to the insanity defense; as the Court stated, "The preponderance of the evidence standard comports with due process for commitment of insanity acquittees" (p. 368). In addition, the Court determined that the defendant's commitment length need not be predetermined by the maximum penalty if convicted, stating,

> There simply is no necessary correlation between severity of the offense and length of time necessary for recovery. The length of the acquittee's hypothetical criminal sentence therefore is irrelevant to the purposes of his commitment. (p. 369)

2.5.3 Mental Illness and Violence Link

Because the correctional system placed such importance upon predicting future dangerousness (e.g., see Beis, 1983, for a state-by-state listing of involuntary commitment statutes; see Monahan, 2006, for a thorough review of court cases), it was clear that alternative methods were needed to do so. However, little research at the time (ca. 1970's) was devoted to successfully predicting dangerousness of the mentally ill; the research that existed yielded unsatisfactory results. As Dix (1976) states,

> [T]here is both need and justification for additional research concerning prediction of the behavior of the mentally ill, even if these studies involve high risks. More studies are needed utilizing carefully matched groups of persons whose commitment has been sought, who have been "determined" to be dangerous, and who refuse to voluntarily submit to treatment. Persons in one group must be subjected to involuntary treatment, while those in the other must simply be left to their own devices. (p. 332)

Although Dix's proposal may be extreme, his point is clear: more adequate research was necessary. An unfortunate common belief is that dangerous behavior and mental illness often coexist; Bauer (1970) went so far as to proclaim that "[s]chizophrenia and criminality (or delinquency) appear to be very closely related—two sides of the same coin" (p. 158). Cocozza, Melick, and Steadman (1978) attempt to debunk this belief; looking at the rates of crime among previously-released patients, they found that a small number were involved in violent crimes and that little violent crimes involved those who are mentally ill. They also found individuals who were younger and had a history of criminal behavior were the most likely to be arrested for violent crimes.

In reanalyzing data from Operation Baxstrom, Cocozza and Steadman (1974) were able to significantly differentiate dangerous patients from nondangerous ones based on age (age 50 or older were more likely to be dangerous) and their score on the Legal Dangerousness Scale (a score of 5 or more indicated a higher likelihood of dangerousness). The Legal

Dangerousness Scale (LDS) was developed by the authors and only consists of four aspects of a patient's previous criminal activity (juvenile record, number of previous arrests, conviction of violent crimes, and severity of offense that led to their most recent incarceration).

Stuart (2003) looks at the relationship between violence and mental illness, concluding that "mental disorders are neither necessary, nor sufficient causes of violence" (p. 123) and that socio-demographic and socio-economic factors and substance abuse were major determinants. Stuart also suggested that the strength of the relationship between mental illness and violence is exaggerated. Quinsey (1979) looked at several types of prediction methods of future dangerousness of mentally-ill patients based on clinical assessments, demographic data, psychometric assessments (e.g., the MMPI), laboratory operant studies, and behavior while institutionalized; he found that only age and diagnosis were important variables in predicting dangerousness.

Swanson, Holzer III, Ganju, and Jono (1990) attempt to clarify the link between mental illness and violence; in their research they analyzed responses of over 10,000 people from Baltimore, Raleigh-Durham, and Los Angeles. The authors report that 368 (3.7%) people admitted behaving violently within the previous year; in addition, they note that more than half of those violent (55.5%) met a DSM-III criteria for a psychiatric disorder (about 20% met these criteria in the overall sample). The authors reported that the risk of violence increased with the number of psychiatric diagnoses and that those with a substance abuse problem or dependence had the highest rate of violence (the next highest rate was among individuals diagnosed as schizophrenic). Finally, the authors found an interaction effect between substance abuse and mental illness.

E. Silver (2006) notes that research suggests that most mentally-ill individuals are not violent but that the likelihood of violence is greater among the mentally ill and that the risk of violence is greatly increased with substance abuse. In their meta-analysis, Douglas, Guy, and Hart (2009) found that psychosis and violence were associated with each other; however, they noted the effect is small with a lot of variability. Elbogen and Johnson (2009)

found that violence was associated with severe mental illness (schizophrenia, bipolar, and major depression), but only when co-occurring with substance abuse or dependence. Results from S. Yang, Mulvey, Loughran, and Hanusa (2012) suggest that alcohol use, recent past violence, and levels of affective symptoms, as measured by the Brief Psychiatric Rating Scale (BPRS; Overall & Gorham, 1962), are associated with violence among depressed patients, but not those with a psychotic disorder.

The (TAC; Treatment Advocacy Center, 2013) lists on its website "preventable tragedies." Their database consists of violent incidences since 1987 that involved individuals with a "neurological brain disorder" that is often untreated. According to their website, under the title "Fast Facts," about 10% of homicides in the United States every year (well over 1,000 a year) are committed by someone with a severe mental illness. One of the "solutions" suggested by the TAC to prevent these crimes is the implementation of involuntary outpatient commitment. In fact, Dr. E. Fuller Torrey and the TAC were largely influential in the development of New York State's so-called "Kendra's Law" (N.Y. Mental Hygiene Law, 1999), which allows the state to force psychiatric treatment upon people meeting specified criteria. The TAC website claims that two different studies "proved" that Kendra's Law helps the mentally ill, protects the public, and saves money.

2.5.4 *Barefoot v. Estelle* (1983)

The early 1980s produced two landmark cases regarding prediction of future dangerousness. In *Estelle v. Smith* (1981), the United States Supreme Court ruled that a forced psychiatric evaluation intended for sentencing was a violation of the Fifth and Sixth Amendments. Two years later, in *Barefoot v. Estelle* (1983), the United States Supreme Court ruled that psychiatrists' predictions of future dangerousness were admissible. As previously mentioned, the American Psychiatric Association opposed the admissibility; the Court's opinion, however, differed:

> There is no merit to petitioner's argument that psychiatrists, individually and as a group, are incompetent to predict with an acceptable degree of reliability that a particular criminal will commit other crimes in the future, and so represent a danger to the community. To accept such an argument would call into question predictions of future behavior that are constantly made in other contexts. ... Nor, despite the view of the American Psychiatric Association supporting petitioner's view, is there any convincing evidence that such testimony is almost entirely unreliable, and that the factfinder and the adversary system will not be competent to uncover, recognize, and take due account of its shortcomings. (*Barefoot v. Estelle*, 1983, Syllabus, p. 882)

J. D. Bloom and Rogers (1987) advise that psychiatrists' opinions in the legal setting be limited, suggesting that psychiatry remain "only within the ethics of consultation and the limitations of knowledge in [psychiatrists'] own field (Bloom & Bloom, 1985)" (p. 852). They also advise against long-term predictions of dangerousness:

> There are viable sentencing roles for psychiatrists that speak to issues of mitigation (Bloom & Bloom, 1982) and rehabilitation without entering into the arena of long-term prediction of dangerousness (p. 852)

Loftus and Monahan (1980) state, "The psychologist who is asked to present expert testimony must face a variety of ethical and other considerations" (p. 276). They present in detail five such considerations: trustworthiness of research findings, generalizability of research findings, presenting both sides when there is no consensus, the probabilistic nature of research findings, and the psychologist's own personal values. This latter consideration is particularly interesting; the authors provide several anecdotes regarding personal experiences as expert witnesses. For instance,

> Knowing how difficult it is to obtain convictions in rape cases, and knowing that the expert testimony is likely to help the defendant by reducing his

> chances of conviction, E. L. is faced with a personal dilemma. In this particular example, the personal dilemma was resolved when she reasoned that although obtaining convictions of actual rapists is of crucial importance, avoiding the mistaken conviction of innocent people is equally important.

An additional anecdote recalls a case involving the death penalty:

> [The death penalty] was predicated on an accurate prediction of violent behavior—a prediction that [J. M.] believed to be scientifically impossible to make. Since he was morally opposed to the death penalty, the testimony presented no moral problem. If the research findings were otherwise, however—if some new technique allowed violent behavior to be more accurately predicted—he would have refused to testify, believing immoral ends to be no less immoral when pursued with scientific means.

The authors conclude, "The only thing worse than the law's uncritical acceptance of psychology would be psychologists' uncritical rejection of the law as an arena worthy of their participation" (pp. 281–282).

2.6 A Second Generation of Violence Prediction

John Monahan may be the most influential person in the area of violence prediction and assessment. Monahan emerged in the 1970s and continues to publish research in the field. Monahan is the primary investigator in the MacArthur Violence Risk Assessment Study (Monahan et al., 2001) that led to the development of the actuarial instrument, the Classification of Violence Risk. In his 1981 book entitled *Predicting Violent Behavior*, Monahan discusses, among other topics, the issues surrounding future violence predictions and statistical techniques for improving clinical predictions. With respect to accuracy of predictions, he states,

> No one insists that prediction be perfect. We do not, after all, require absolute certainty for convicting the guilty, only proof beyond a "reasonable doubt." This

> means that we are willing to tolerate the conviction of a few innocent persons to assure the confinement of a much larger number of guilty criminals. It also means that, when there is doubt, we would much rather release a guilty person than confine an innocent one. (p. 34)

In noting the different consequences associated with predictions (e.g., denial of parole versus a death penalty sentence), he says, "Far too often we treat predictions as if they were cheap socks: 'One size fits all'" (p. 40). With respect to actuarial methods, Monahan suggested that clinicians emphasize base rates for violence, consider information for valid predictive relationships, and do not overreact to positive associations. With respect to violence and the mentally ill, he states,

> Mental illness ... does not appear to be related to violence in the absence of a history of violent behavior. When one controls for demographic variables, prisoners do not appear to have a higher incidence of mental illness than the general population. Mental patients who do not have a record of violent arrests are, if anything *less* [emphasis added] violent than the general population. (p. 127)

In an article that was originally presented at the Annual Convention of the American Psychological Association in 1991, Monahan (1992) again discusses the relationship between mental illness and violence. Although he previously held the belief that the relationship was non-existent, he states,

> I now believe that there may be a relationship ..., one that cannot be fobbed off as chance or explained away by other factors that may cause them both. The relationship, if it exists, probably is not large, but may be important both for legal theory and for social policy (Monahan, 1992, p. 511).

2.6.1 Rise of the Machines

As computing ability became more powerful, analyzing large data sets using sophisticated techniques became feasible. In his book *Mathematical Criminology*, Greenberg (1979)

presents numerous quantitative methods for analyzing and predicting crime; these include multivariate statistical methods (e.g., path analysis, time series analysis, principal components analysis, factor analysis, discriminant function analysis, cluster analysis, and multidimensional scaling), probability theory and stochastic processes (e.g., the use of Bayes' Theorem, Markov Chains, Poisson processes), and analytical methods (differential equations and Laplace transforms).

Holland et al. (1983) use multiple discriminant analysis to predict violent and nonviolent recidivism using only prior violent and nonviolent convictions as separate predictors and noting an attenuating effect of age, but with poor results; the authors conclude, "Violence is once again found to be minimally predictable" (p. 181). E. Silver, Smith, and Banks (2000) compared iterative schemes (e.g., iterative classification trees) with non-iterative ones and concluded that the iterative schemes were superior in terms of correct classifications. Neuilly, Zgoba, Tita, and Lee (2011) use decision tree methods to predict recidivism among homicide offenders released on parole and found these models performed well, outperforming logistic regression. Tangney, Stuewig, and Martinez (2014) use structural equation models to show that shame is a direct effect of recidivism.

2.6.2 Second Generation Risk Assessment Instruments

Monahan (1984) declares that violence-prediction research was entering a second generation, emphasizing the limitations and mostly disappointing results in the so-called first generation. He notes the second generation of violence prediction would be marked by improved predictive technology, calling for a focus on actuarial methods of prediction that incorporate clinical information. In a different article, Monahan (1988) states,

> To overcome the problems that have so far hobbled the scientific study of violence among the mentally disordered, we must enrich our predictor variables, strengthen our criterion variables, exploit natural variation in validation samples, and synchronize our research efforts. If we do, it is possible that the next generation of risk

> assessment studies will yield results quite different than those to which we have become accustomed. (p. 255)

The "second generation" of violence prediction indeed provided more research using actuarial devices. The last two decades of the 20th century and the early 21st century has seen the development of numerous prediction tools, including, among many others, the Statistical Information on Recidivism (SIR; Nuffield, 1982), developed to predict criminal recidivism; the Classification of Violence Risk (COVR; Monahan et al., 2001; Steadman et al., 2000; Monahan et al., 2000), an actuarial tool based on an iterative classification tree method and used to assess dangerousness in mentally-ill patients; the Violence Risk Appraisal Guide (VRAG; G. T. Harris, Rice, & Quinsey, 1993; C. Webster, Harris, Rice, Cormier, & Quinsey, 1994; Quinsey, Harris, Rice, & Cormier, 2006) and its revised version (VRAG-R; M. E. Rice, Harris, & Lang, 2013), developed using multiple discriminant analysis and used for predicting violent and sexual recidivism; the Psychopathy Checklist (PCL; Hare, 1980), the Psychopathy Checklist–Revised (PCL-R; Hare, 1991), the Psychopathy Checklist: Screening Version (PCL:YV; S. D. Hart, Cox, & Hare, 1995), and the Psychopathy Checklist: Youth Version (PCL:YV; Forth, Kosson, & Hare, 2003), all developed with principal component analysis and used for diagnosing psychopaths (not an actuarial assessment per se, but a psychological one often associated with dangerous behavior); the Level of Service (Supervision) Inventory (Andrews, 1988, LSI) and its revised edition (LSI-R; Andrews & Bonta, 1995) and screening version (LSI-SV; Andrews & Bonta, 1998), designed to predict parole outcome and recidivism; the Historical Clinical Risk–20 (HCR-20; C. D. Webster, Eaves, Douglas, & Wintrup, 1995; Version 2, C. D. Webster, Douglas, Eaves, & Hart, 1997), consisting of ten historical, five clinical, and five risk variables used for risk assessment; the Violence Screening Checklist (VSC; McNiel & Binder, 1994a; McNiel & Binder, 1994b), a five-item actuarial tool used to assess risk of future violence and the Violence Screening Checklist–Revised (VSC-R; McNiel, Gregory, Lam, Binder, & Sullivan, 2003, the VSC with the fifth item omitted); the Sexual Violence Risk–20 (SVR-20; Boer, Hart, Kropp, & Webster, 1997), an assessment instrument

used to determine an individual's risk of committing sexual violence; the Rapid Risk Assessment for Sex Offense Recidivism (RRASOR; Hanson, 1997), used to screen sex offenders to estimate risk of recidivism; the Structured Anchored Clinical Judgement scale (SAC-J; Grubin, 1998); the Static-99 (Hanson & Thornton, 2000), the Static-2002 (Hanson & Karl, 2003), and their revised versions (Static-99R; Static-2002R; Helmus, Thornton, et al., 2012), all being prediction measures for long-term risk of sexual and violent recidivism for male sex offenders; the Minnesota Sex Offender Screening Tool (MnSOST; Epperson, Kaul, & Huot, 1995) and its revised versions (MnSOST-R, Epperson, Kaul, & Hesselton, 1998; MnSOST-3, Duwe & Freske, 2012), developed to estimate the risk of sexual recidivism in male sex offenders; the Sexual Offender Risk Appraisal Guide (SORAG; Quinsey, Rice, & Harris, 1995), an actuarial tool for predicting sexual recidivism among sex offenders; the Domestic Violence Risk Appraisal Guide (DVRAG; Hilton, Harris, Rice, Houghton, & Eke, 2008), an actuarial instrument assessing the risk of recidivism among (male) domestic assaulters; the Spousal Assault Risk Assessment Guide (SARA; Kropp, Hart, Webster, & Eaves, 1994; Kropp, Hart, Webster, & Eaves, 1999; Kropp & Hart, 2000), a structured professional judgment (SPJ) tool consisting of twenty items for assessing risk of violence in domestic violent perpetrators and developed for use in the criminal justice system; the Short-Term Assessment of Risk and Treatability (START; C. D. Webster, Martin, Brink, Nicholls, & Middleton, 2004; C. D. Webster, Nicholls, Martin, Desmarais, & Brink, 2006), a SPJ tool for informing risk in several domains such as suicide, substance abuse, and violence toward others; the Offender Group Reconviction Scale (OGRS; Copas & Marshall, 1998) and its revised versions (OGRS-2, R. Taylor, 1999; OGRS-2, Howard, Francis, Soothill, & Humphreys, 2009), a twelve-item assessment tool for measuring risk of reoffense and harm; the Risk Matrix 2000 for Violence (RM2000V; Thornton, 2007), a three-item rating instrument for predicting nonsexual violence in adult males; the Lifestyle Criminality Screening Form (LCSF; Walters, White, & Denney, 1991), a fourteen-item screening instrument for identifying lifestyle criminality; the Violence Risk Scale (VRS; Wong & Gordon, 1999) and the Violence

Risk Scale–Sexual Offender version (VRS-SO; Olver, 2003; Olver, Wong, Nicholaichuk, & Gordon, 2007), actuarial instruments for predicting (sexual and nonsexual) violence that also measures treatment changes using the "Stages of Change" model; the Risk for Sexual Violence Protocol (RSVP; S. D. Hart, Kropp, & Laws, 2003), an SPJ for assessing sexual violence risk; the Risk Assessment Scale for Prison (RASP; Cunningham & Sorensen, 2006, 2007), a scale for assessing violence risk among prisoners; the Early Assessment Risk List for boys (EARL-20B; Augimeri, Webster, Koegl, & Levene, 1998 Augimeri, Koegl, Webster, & Levene, 2001) and the Early Assessment Risk List for girls (EARL-21G; Version 1, Levene et al., 2001), gender-specific SPJ risk assessment tools for predicting juvenile delinquency; the Structured Assessment of Violence Risk in Youth (SAVRY; Bartel, Forth, & Borum, 2003; Borum, Bartel, & Forth, 2006; Borum, Bartel, & Forth, 2005), designed to measure the violence risk in adolescents (ages 12–18); and the Youth Level of Service/Case Management Inventory (YLS/CMI; Hoge & Andrews, 2002), an actuarial risk/need assessment tool designed to evaluate the risk and needs of troubled youth.

Many studies validating these instruments followed. For example, Grann, Belfrage, and Tengström (2000) assessed the validity of the HCR-20 and VRAG in a mentally-ill Swedish population. Monahan et al. (2005) validated the COVR in an independent sample and Sturup, Kristiansson, and Lindqvist (2011) found the COVR predicted violence in a Swedish population. Dolan and Doyle (2000) discuss the utility of the PCL, PCL-R, and PCL:SV in predicting violence. Barbaree, Seto, Langton, and Peacock (2001) evaluated the accuracy of the VRAG, SORAG, RRASOR, Static-99, MnSOST-R, and PCL-R for predicting general, sexual, or serious (sexual and violent) recidivism. Buffington-Vollum, Edens, Johnson, and Johnson (2002) studied the relationship between psychopathy (using the PCL-R) and serious institutional misconduct in an incarcerated sex offender population. Doyle, Dolan, and McGovern (2002) examined the validity of the PCL:SV, HCR-20, and VRAG for predicting inpatient violence in an English hospital for the mentally ill. N. S. Gray et al. (2003) looked at the efficacy of the HCR-20 and PCL-R, among others, for predicting self-

harm and violence toward others. G. T. Harris et al. (2003) compared the VRAG, SORAG, RRASOR, and Static-99 in predicting violent and sexual recidivism among four different samples of sex offenders. Nicholls, Ogloff, and Douglas (2004) looked at the predictive validity for violence risk assessment among the HCR-20, the PCL:SV, the VSC in a sample of involuntarily hospitalized males and female forensic patients. Worth noting is that one of the items of the VRAG is the PCL; Edens, Skeem, and Douglas (2006) examined the *incremental validity* (i.e., a measure's improved predictive accuracy over another measure) of the VRAG items (excluding the PCL score) to the PCL:SV and conclude that "the validity of the modified VRAG was attributable primarily to 1 of its 10 items—the PCL:SV" (p. 371).

In addition, the second generation included several meta-analytic studies (although some of these included research from the first generation as well). For example, Bonta, Law, and Hanson (1998) conducted a meta-analysis on 58 studies from the years 1959–1995 and found that predictors of recidivism among mentally-ill offenders were no different than those in nondisordered ones. They also report that criminal-history variables displayed far larger effect sizes than clinical variables. Hanson and Morton-Bourgon (2007) looked at 79 different samples examining the predictive accuracy of actuarial, SPJ, and clinical predictions for sexual, violent, or general recidivism; they found that, on average, actuarial measures perform the best with SPJ next and clinicians the worst, but that the best overall measure was the SVR-20, an SPJ device. Helmus, Hanson, Thornton, Babchishin, and Harris (2012) found the Static-99R and Static-2002R to be "remarkably consistent" across the 23 samples of sex offenders that they examined; they also found that the rate of sexual recidivism is quite lower than most would believe and what previous research has shown. In one of the most exhaustive meta-analyses in the research area, Singh, Grann, and Fazel (2011) examined nine risk assessment measures (LSI-R, PCL-R, SORAG, Static-99, VRAG, HCR-20, SVR-20, SARA, and SAVRY) from sixty-eight studies consisting of eighty-eight independent samples totaling 25,980 individuals from thirteen countries. The authors concluded the predictive validity of the measures differed significantly; they found that the SAVRY performs the best, whereas

the LSI-R and PCL-R performs worst. The authors suggest a reason the SAVRY may have done so well is because it is used on populations it was designed to be for (i.e., adolescent offenders). In a separate article, Fazel, Singh, Doll, and Grann (2012) looked at the same nine risk assessment measures and found that the measures produce high sensitivities (median for violent offenses: .92; for sexual offenses: .88) but low positive predictive values (median for violent offenses: .41; for sexual offenses: .23). In another exhaustive meta-analysis examining nine risk assessment instruments used for predicting violence (PCL-R, PCL:SV, VRAG, HCR-20, LSI-R, OGRS, SIR, VRS, and RM2000V), M. Yang, Wong, and Coid (2010) conclude that all nine instruments predicted violent recidivism "moderately well" and they were not significantly different from one another. The authors also advise that "[b]ecause of their moderate level of predictive efficacy, [the nine instruments] should not be used as the sole or primary means for clinical or criminal justice decision making that is contingent on a high level of predictive accuracy, such as preventive detention" (p. 761).

2.6.3 The Prisoner Cohort Study

The Prisoner Cohort Study (PCS; Coid et al., 2007) was one of the largest studies in the field consisting of 1470 prisoners in the United Kingdom were interviewed; all prisoners were serving a sentence that was a minimum of two years for a violent or sexual offense and were to be released within a year. All prisoners were over the age of eighteen (mean age 30.8; maximum age 75); 95% ($N = 1396$) were male; 79% were white, 15% black, 3% Asian, and 3% were of another ethnic origin. Several predictive measures were used on the sample: the VRAG, the HCR-20, RM2000V, and the OGRS-2 as well as the PCL-R. The male sample was the sample used in the analyses. Follow-up time was within three years after being released from prison. Of the 1396 male prisoners, 43 were not released in this time period; 41% of the sample were reconvicted.

The predictive accuracy of the OGRS-2 was highest (as measured by AUC) for all types of reconvictions (violence, robbery, theft, drugs, and any), but all measures were

significant. In addition, the researchers were interested in prisoners with dangerous and severe personality disorders (defined as having at least two personality disorders and a PCL-R score less than 25); at least one personality disorder (excluding antisocial personality disorder) and a PCL-R score between 25-29; or a PCL-R score of greater than 30); 212 prisoners (15%) fit the criteria. The DSPD prisoners who were released accounted for 27% of all reconvictions, 25% if violent crimes, and 0% of the sexual offenses (recall that the overall reconviction rate was 41%; for violent offenses it was 11.5% and for sexual offenses it was only 0.5%). The authors concluded that the probability of reconviction was significantly higher for DSPD prisoners than non-DSPD prisoners for any and violent offenses.

Using data from the PCS, M. Yang, Liu, and Coid (2010) compared the predictive power of several classification methods: decision trees, neural networks, logistic regression, and discriminant analysis. The authors found that the four methods performed comparably. When predicting violent recidivism versus no recidivism (i.e., excluding nonviolent recidivism as an outcome), their models performed well.

2.7 Sexually Violent Predators

Sexually violent predators[3] are some of the most despised criminals; nonetheless. As District Judge Terrence Boyle stated,

> Sexual predators and child molesters are among the most villainous in our society, and the government has a duty—within the bounds of the Constitution—to protect its populace from these those people who are likely to harm others. Given the often abhorrent nature of these individuals' criminal backgrounds, courts may be tempted to turn a blind eye to any due process violations. But the Courts have a duty to protect the rights of even the most despised among us. (*United States v. Edwards*, 2011, pp. 996–997)

[3]A majority of states use this term, but some use less provocative terms such as sexually violent persons or sexually dangerous persons/individuals (see Deming, 2008)

By the 1950s "an increasingly large number of ... states in the United States[] ha[d] passed special statutes ... dealing with so-called 'psychopathic' sexual offenders" (Hacker & Frym, 1955, p. 766), commonly referred to as *Sexual Psychopath Acts* (SPA; for a historical review of sexual psychopath laws, see Lave, 2008).

Under Washington D.C.'s Sexual Psychopath Act, a man named Maurice Millard was committed until "restored to mental competence" (*Millard v. Harris*, 1968); in *Millard v. Harris* (1968), Millard attacked the constitutionality of the D.C. SPA. Providing his opinion to the Court, Chief Judge David Bazelon discussed civil commitment based on dangerousness, stating,

> [C]onstitutional issues of the gravest magnitude immediately appear. Substantively, there is serious question whether the state can ever confine a citizen against his will simply because he is likely to be dangerous in the future, as opposed to having actually been dangerous in the past. ... Predictions of dangerousness, whether under the Sexual Psychopath Act or in some other context, require determinations of several sorts: the type of conduct in which the individual may engage; the likelihood or probability that he will in fact indulge in that conduct; and the effect such conduct if engaged in will have on others. Depending on the sort of conduct and effect feared, these variables may also require further refinement. Our evaluation of the ultimate dangerousness of certain forms of behavior may vary with the frequency with which they can be expected. If so, it will be necessary to evaluate not only the likelihood that the individual will misbehave in such fashion, but also the probability that he will offend with a certain frequency. And since the effect on others may depend on who the victim is, an estimate of the likelihood that a certain sort of person may prove the victim may also be necessary. ... [A]n an examination of all aspects of the problem is essential. (*Millard v. Harris*, 1968, p. 973)

In *Cross v. Harris* (1969), Chief Judge Bazelon again gave his opinion regarding the issue:

> To be "dangerous" for the purposes of the Sexual Psychopath Act, one must be likely to attack or otherwise inflict injury, loss, pain, or other evil on the objects of his desire. The focus of the statute is not on expected conduct, but on the harm that may flow from that conduct. Commitment cannot be based simply on the determination that a person is likely to engage in particular acts. The Court must also determine the harm, if any, that is likely to flow from these acts. A mere possibility of injury is not enough; the statute requires that the harm be likely. For no matter how certain one can be that a person will engage in particular acts, it cannot be said that he is "likely to inflict injury" unless it can also be said that the acts, if engaged in, are likely to result in injury. (pp. 1099–1100)

2.7.1 Sexually Violent Predator Laws

In 1990, Washington State was the first to establish a sexually violent predator (SVP) law that allowed sex offenders to be civilly committed—even after the completion of a sentence—if the person was deemed a sexually violent predator. As of 2014, nineteen more states (Arizona, California, Florida, Illinois, Iowa, Kansas, Massachusetts, Minnesota, Missouri, Nebraska, New Hampshire, New Jersey, New York, North Dakota, Pennsylvania, South Carolina, Texas, Virginia, and Wisconsin), along with the District of Columbia, have enacted similar laws (The Association for the Treatment of Sexual Abusers, 2014).

Chapter 980 of the Wisconsin Statute (Wis. Stats. ch. 980, 2013) outlines the civil commitments of sexually-violent persons for the State of Wisconsin, defining a "sexually violent person" as follows:

> [A] person who has been convicted of a sexually violent offense, has been adjudicated delinquent for a sexually violent offense, or has been found not guilty of or not responsible for a sexually violent offense by reason of insanity or mental disease, defect, or illness, and who is dangerous because he or she suffers from a mental disorder that makes it likely that the person will engage in one or more acts of

sexual violence. (§.01(7))

When the Court finds an individual to be sexually violent, it may order that person to be committed until "no longer a sexually violent person" (§.06). In *State v. Post* (1995), the Supreme Court of Wisconsin held that the Chapter 980 does not violate the Constitutions of Wisconsin or the United States.

Kansas v. Hendricks **(1997)**

The State of Kansas defines a sexually violent predator as one "who has been convicted of or charged with a sexually violent offense and who suffers from a mental abnormality or personality disorder which makes the person likely to engage in repeat acts of sexual violence" (Kan. Stats. ch. 59, Art. 29a, 2012, §02(a)). The statute defines "mental abnormality" as the "congenital or acquired condition affecting the emotional or volitional capacity which predisposes the person to commit sexually violent offenses in a degree constituting such person a menace to the health and safety of others" (§02(b)); it defines "likely to engage in repeat acts of sexual violence" as "the person's propensity to commit acts of sexual violence is of such a degree as to pose a menace to the health and safety of others" (§02(c)). The Sexually Violent Predator (SVP) Act of Kansas, enacted in 1994, allowed for indefinite civil commitment for a convicted sex offender deemed a sexually-violent predator. The first offender committed under the act was Leroy Hendricks, who was scheduled to be released shortly after the act was passed. Hendricks challenged the constitutionality of the commitment, specifically that it violated due process, double jeopardy, and ex post facto clauses. The case reached the United States Supreme Court in *Kansas v. Hendricks* (1997) and was the first SVP law to be challenged in the highest court. The Supreme Court ruled in favor of the act 5–4.

Justice Clarence Thomas gave the majority opinion, with Justices Antonin Scalia, Sandra Day O'Connor, Anthony Kennedy, and William Rehnquist joining. According to Thomas,

> A finding of dangerousness, standing alone, is ordinarily not a sufficient ground upon which to justify indefinite involuntary commitment. We have sustained civil commitment statutes when they have coupled proof of dangerousness with the proof of some additional factor, such as a "mental illness" or "mental abnormality." These added statutory requirements serve to limit involuntary civil confinement to those who suffer from a volitional impairment rendering them dangerous *beyond their control* [emphasis added]. The Kansas Act is plainly of a kind with these other civil commitment statutes: It requires a finding of future dangerousness, and then links that finding to the existence of a "mental abnormality" or "personality disorder" that makes it difficult, if not impossible, for the person to control his dangerous behavior. The precommitment requirement of a "mental abnormality" or "personality disorder" is consistent with the requirements of these other statutes that we have upheld in that it narrows the class of persons eligible for confinement to those who are unable to control their dangerousness. (p. 358, citations omitted)

Justice Kennedy, in his concurring statement, warned,

> [C]ivil confinement were to become a mechanism for retribution or general deterrence, or if it were shown that mental abnormality is too imprecise a category to offer a solid basis for concluding that civil detention is justified, our precedents would not suffice to validate it. (p. 373)

Justice Stephen Breyer wrote the dissent, with Justices David Souter, Ruth Ginsburg, and John Stevens joining (note that Justice Ginsburg only joined on two of the three parts). In the dissent, Breyer agreed with the Court that the Kansas law did not violate the due process clause of the constitution (Justice Ginsburg did not join this part), but argued that it violated the double jeopardy and ex post facto clauses. In particular, the feeling among the dissenting judges was that the commitment of Hendricks was punitive, rather than civil, and meant to further punish him.

The Kansas statute appeared again in the United States Supreme Court in *Kansas v. Crane* (2002), where the Court addressed the "emotional or volitional capacity" clause in the Kansas statute. Particularly, the state of Kansas felt that the "Kansas Supreme Court wrongly read *Hendricks* as requiring the State *always* [original emphasis] to prove that a dangerous individual is *completely* [original emphasis] unable to control his behavior" (p. 411), to which the United States Supreme Court replied, "We do not agree with the State, however, insofar as it seeks to claim that the Constitution permits commitment of the type of dangerous sexual offender considered in *Hendricks* without *any* [original emphasis] lack-of-control determination" (p. 412). The Court continued,

> [In *Hendricks*,] we did not give to the phrase "lack of control" a particularly narrow or technical meaning. And we recognize that in cases where lack of control is at issue, "inability to control behavior" will not be demonstrable with mathematical precision. It is enough to say that there must be proof of serious difficulty in controlling behavior. (p. 413)

As Faigman et al. (2013) point out, in *Kansas v. Crane* the Supreme Court essentially determined that "psychologists and psychiatrists can distinguish between mentally *abnormal* [original emphasis] and dangerous sex offenders and mentally *normal* [original emphasis] and dangerous sex offenders" (p. 182).

The Court ruled 7–2, with Justice Scalia dissenting and Justice Thomas joining. Scalia stated,

> There is good reason why, as the Court accurately says, "when considering civil commitment . . . we [have not] ordinarily distinguished for constitutional purposes among volitional, emotional, and cognitive impairments." We have not done so because it makes no sense. It is obvious that a person may be able to exercise volition and yet be unfit to turn loose upon society. The man who has a will of steel, but who delusionally believes that every woman he meets is inviting crude sexual advances, is surely a dangerous sexual predator. (p. 422)

He continued,

> I suspect that the reason the Court avoids any elaboration is that elaboration which passes the laugh test is impossible. How is one to frame for a jury the degree of "inability to control" which, in the particular case, "the nature of the psychiatric diagnosis, and the severity of the mental abnormality" require? (p. 423)

Scalia, in his concluding paragraph, suggests that when an individual is declared mentally ill, it means that he is unable to control his behavior:

> A jury determined beyond a reasonable doubt that respondent suffers from antisocial personality disorder combined with exhibitionism, and that this is either a mental abnormality or a personality disorder making it likely he will commit repeat acts of sexual violence. That is all the SVP[Act] requires, and all the Constitution demands. (p. 425)

Despite the Supreme Court's ruling, it has not been followed by many lower-level courts; several states have adopted Scalia's suggestion that mental illness and dangerousness together imply lack of control (Pierson, 2011). For a thorough review and the implications of the two cases, see Faigman et al. (2013).

Adam Walsh Child Protection Act of 2006

In New Jersey, Megan's Law requires all sex offenders to register with the state; the law characterizes sex offenders along three risk levels:

> The regulations shall provide for three levels of notification depending upon the risk of re-offense by the offender as follows:
> (1) If risk of re-offense is low, law enforcement agencies likely to encounter the person registered shall be notified;
> (2) If risk of re-offense is moderate, organizations in the community including schools, religious and youth organizations shall be notified in accordance with the

> Attorney General's guidelines, in addition to the notice required by paragraph (1) of this subsection;
>
> (3) If risk of re-offense is high, the public shall be notified through means in accordance with the Attorney General's guidelines designed to reach members of the public likely to encounter the person registered, in addition to the notice required by paragraphs (1) and (2) of this subsection. (New Jersey Code of Criminal Justice, 2014, 8-3.(c))

As explained by the New Jersey Supreme Court in *Doe v. Poritz* (1995),

> We are aware of the uncertainties that surround all aspects of the subject of sex offender recidivism and the effectiveness of preventive measures. Legislatures, despite uncertainty, must sometimes act to deal with public needs, basing such action on what they conclude, in a welter of conflicting opinions, to be the probable best course. Our Legislature could reasonably conclude that risk of reoffense can be fairly measured, and that knowledge of the presence of offenders provides increased defense against them. Given those conclusions, the system devised by the Legislature is appropriately designed to achieve the laws' purpose of protecting the public

In July 2006, then-U.S. president George W. Bush signed into law the Adam Walsh Child Protection Act of 2006 (AWA; H.R. 4472, 2006). The act was written "[t]o protect children from sexual exploitation and violent crime, to prevent child abuse and child pornography, to promote Internet safety, and to honor the memory of Adam Walsh[4] and other child crime victims" (p. 1). Under Section 302, the AWA permits the federal authority to civilly commit an SVP for a possibly indefinite period of time, which is often the case (Prescott, 2009). In *United States v. Comstock* (2010), the United States Supreme Court determined the Federal Government has the authority to civilly commit individuals in federal custody as allowed by the AWA.

[4]Adam Walsh was a six-year-old Florida boy who was abducted and murdered in 1981.

By 2002, Fitch (2003) reports that nearly 2500 SVPs were committed or pending a commitment hearing in the United States; in states that allowed for SVP commitments to move to outpatient treatment, there were only 69 such reported cases. He also notes that "[Eighty-two] people committed as SVPs (nationally) had been released from confinement" (p. 492). According to Deming (2008), by 2006 the number of civilly committed SVPs has increased to over 3600.

2.7.2 "Likelihood" of Dangerousness

The Kansas statute—and many others like it—require that a person must be found dangerous to be committed. Examining the different statutes for SVP commitment laws leads one to note the differing risk thresholds required for commitment. Depending on the state, the standard for commitment is usually defined as "likely," or something similar. Some states have formally defined standards for commitment; for example, Washington State defines "likely to engage in predatory acts of sexual violence if not confined in a secure facility" as the person that "more probably than not will engage in such acts" (Revised Code of Washington, 2013; see also *In re Brooks*, Wash. 2001 or *In re Hosier*, Wash. 2010). The Supreme Court of Wisconsin decided that the term used in their statute (Wis. Stats. ch. 980, 2013), "substantially probable," meant "much more likely than not" (*State v. Curiel*, 1999).

Sreenivasan, Weinberger, and Garrick (2003) note the differences or vagueness in state definitions and suggests these definitions "are best conceptualized through a systematic and structured approach to risk assessment in which the explanation for the opinion is clear" (p. 484) and give suggestions for doing so. But few states have quantified these terms (e.g., see *Doe v. Poritz*, 1995).

In their report comparing the different state SVP laws (there were only 12 in 1998), Lieb and Matson (1998) examined the differences between the processes, decision steps, and costs. They report that most states detain SVPs for an indeterminate time (e.g., Arizona);

that some states require proof beyond a reasonable doubt (e.g., Iowa), whereas others require only clear and convincing evidence (e.g., Florida); and that some SVP laws are applicable to juveniles (e.g., Illinois). One of the more relevant details of their report considers the differing likelihood standards set by the courts to civilly commit an individual using previously discussed terms such as "likely to engage in" or "substantially probable that the person will engage in" in reference to "acts of sexual violence." They also note that these probabilities are commonly determined using actuarial methods, such as the Static-99.

As determined in *Addington v. Texas* (1979), civil commitment may require a less stringent burden of proof. But as Faigman et al. (2013) point out, treatment is part of the reason for commitment and "[l]aws that commit those who are allegedly sexually dangerous persons promise no similar benefits and are drawn and defended on the bases of specific deterrence and public safety" (p. 216). Nonetheless, the AWA states that if "the court finds by *clear and convincing evidence* [emphasis added] that the person is a sexually dangerous person, the court shall commit the person to the custody of the Attorney General" (H.R. 4472, 2006, § 4248.(d)).

In *United States v. Wooden* (2012), the United States Court of Appeals noted that the district court, who allowed the release of the appellee Walter Wooden, "erred by insisting that the government prove Wooden's 'dangerousness,' which the court believed required proof of a greater-than-50% risk that Wooden would re-offend within five years" (pp. 460–461). The district court concluded,

> [B]ecause none of the actuarial risk-assessments models [used to assess Wooden's risk of recidivism] showed a five-year recidivism rate of 50% or more, 'the actuarial instrument scores alone cannot possibly satisfy the statutory threshold of clear and convincing evidence that Mr. Wooden would have serious difficulty refraining from engaging in sexually violent conduct or child molestation. (p. 450)

The California Supreme Court made a similar statement in *People v. Superior Court (Ghilotti)* (2002), saying that "'*likely* [original emphasis] to engage in acts of sexual violence'

... does not mean the risk of reoffense must be higher than 50 percent" (p. 23). The Court also said,

> An evaluator's conclusion that one does not meet the criteria for commitment or recommitment is legally erroneous if it stems from a conclusion that, although the person presents a serious and well-founded risk of reoffense if free without conditions, the evaluator cannot say the risk exceeds 50 percent. (p. 6)

As Faigman et al. (2013) points out, the Court used a thesaurus, rather than a dictionary, to define "likely" and defend their decision. This was also pointed out in the dissent by Justice Katheryn Werdegar, who notes one dictionary defines likely as "having a better chance of existing or occurring than not" (p. 35).

Additionally, predictions of future dangerous behavior do not dictate the time frame of said behavior (Mossman, Schwartz, & Elam, 2011); in other words, the individual deemed violent may not commit an act of violence in the *immediate* future, as noted in *Purifoy v. Watters* (2010):

> Nowhere in [*Kansas v.*]*Hendricks* or [*Kansas v.*]*Crane* did the Supreme Court rule that there must be an *imminent* [original emphasis] danger presented by an offender in order for him to be civilly committed as a sexually violent person within the meaning of a state statute.

2.7.3 Predicting Sexual Recidivism

In a seven-year follow-up period, Song and Lieb (1995) report in a 1995 study of Washington State sex offenders that the sexual recidivism rate was about 12 percent—43 percent of those rearrest crimes were worse than the original. In addition, they report the average time of rearrest to be 4.6 years. For any felonious crime (sexual, violent, or other), the recidivism rate was 23 percent. The authors note several factors significantly associated with recidivism: age (younger more likely for any type of felonious recidivism), criminal history

(repeat offenders more likely to recidivate for sexual or other felonious crime), race (non-white more likely to recidivate for violent crime), conviction type (rapists more likely to recidivate for sexual or felonious crime), and sentence length (offenders with longer sentences more likely to be arrested for a new sexual offense). Focusing strictly of convicted extrafamilial child molesters, Firestone et al. (2000) found the recidivism rates (after 12 years) were 15%, 20%, and 42% for sexual, violent, or any criminal offense, respectively.

There exists a myriad of research examining the usefulness of risk assessment instruments (RAIs) for predicting recidivism in sex offenders; some support their use, some do not. Gretton, McBride, Hare, O'Shaughnessy, and Kumka (2001) found psychopathy (as measured by the PCL:YV) was positively associated (as measured by correlations) with total, violent, and nonviolent (but not sexual) recidivism in adolescent sex offenders. Using Chi-square tests and odds ratios, they found that offenders in the high psychopathy category were more likely to recidivate than offenders with low psychopathy scores. Hildebrand, De Ruiter, and de Vogel (2004) concluded that the PCL-R was a significant predictor of recidivism among convicted rapists in the Netherlands. Rettenberger, Boer, and Eher (2011) found that the SVR-20 was an accurate predictor of sexual recidivism in an Austrian sample, although they note some inconsistencies. Other studies providing results supporting these measures for predicting sexual recidivism include, among others, the following: VRAG, SORAG, Static-99, and RRASOR across multiple sites (G. T. Harris et al., 2003); PCL-R among rapists (Hildebrand et al., 2004); VRAG, SORAG, RRASOR, Static-99, Static-2002, and MnSOST-R across several previous studies (Langton et al., 2007); VRAG among rapists and child molesters (M. E. Rice & Harris, 1997); RRASOR and Static-99 among Swedish child molesters and rapists (Sjöstedt & Långström, 2001). All studies used AUCs to measure predictive accuracy. Doren (2004b) suggests a multidimensional model for sexual recidivism risk, suggesting two such dimensions might be "sexual deviance" and "criminal personality;" Doren pointed to two different factor analytic studies supporting this idea (Butz-Whittaker, Strassberg, & the Center for Family Development, 2001; C. F. Roberts, Doren, & Thornton,

2002). Allan, Grace, Rutherford, and Hudson (2007) found a four-factor model that was significantly correlated with sexual recidivism among child molesters and, when combined with the Static-99, increased the predictive accuracy.

Bartosh, Garby, Lewis, and Gray (2003) looked at the effectiveness for predicting sexual recidivism of four RAIs, the Static-99, RRASOR, MnSOST-R, and SORAG. The authors found that only the Static-99 and SORAG showed significant predictive validity (based on significant AUCs at a $p < .05$ threshold). Specifically looking at offender types, the authors note that the Static-99 and SORAG showed significant predictive validity with extrafamilial child molesters; the Static-99 also showed significant predictive validity for incest offenders. All four instruments failed to show any predictive validity for sexual recidivism among rapists and "hands-off" offenders. Sjöstedt and Långström (2002) examined the predictive accuracy (based on AUC measures) four risk assessment tools (RRASOR, SVR-20, PCL-R, and VRAG) and found that the RRASOR was the only instrument able to predict sexual recidivism better than chance among rapists. Sjöstedt and Grann (2002) determined the accuracy of two tools for assessing sexual recidivism risk depends largely on the type of recidivism the user is interested in; for instance, they report that the RRASOR and Static-99 do not discriminate intrafamilial and non-intrafamilial recidivists.

Bechman (2001) is particularly critical of the use of RAIs in SVP commitments, stating, "[T]here is increasing support for the position that, not only are [A]RAIs not the 'best science,' they are not science at all" (p. 26). Janus and Prentky (2003) express their sentiments regarding the use of RAIs with SVPs:

> [I]t is logically incoherent to exclude evidence that presumptively improves upon the reliability and accuracy of these judgments ... *if* [original emphasis] courts deprive people of liberty based on assessed risk, then [actuarial]RA[Is] should be part of the assessment. Courts should use [actuarial]RA[Is] in part because it will improve risk assessment. (pp. 1445–1446)

However, the authors caution a sort of catch-22:

> Improved ability to identify persons at high risk for violence may make expanded preventive detention laws politically impossible to resist. New laws, in turn, may demand better risk assessment, which may beget more aggressive and expansive prevention laws, and so on. (p. 1445)

2.8 New Generation, Old Problems

The so-called second generation of research brought along many of the first generation issues. Numerous articles continued to question the ethical implications of predicting future dangerous behavior (e.g., Litwack, 1993). And although many of the second-generation actuarial methods greatly improved upon previous methods and clinical judgment, they were not without their flaws. Corrado (1996), commenting on the notion that for every true positive, there is at least one false positive (based on the then-current research), says, "Th[e] rate of error would seem to many to be unacceptably high; it is much higher, for example, than the rate of error that we would find acceptable in convicting people of crimes" (p. 792).

Grisso and Appelbaum (1992), arguing against the notion that future violent behavior predictions are by definition unethical, cited four distinct types of violence predictions: *dichotomous* (future violent acts will or will not occur), *dichotomous with qualified confidence* (future violent acts will or will not occur, along with the confidence of that happening [e.g., "very likely"]), *risk, individual-based* (the degree of likelihood future violent acts will occur with respect to an individual [e.g., "20% probability person will be violent"]), and *risk, class-based* (degree of likelihood future violent acts will occur with respect to a group the individual is assumed to be a member). As they note, most earlier research (i.e., from the first generation) was in the form of dichotomous predictions, particularly predictions made from clinicians. (Note that probabilistic risk predictions were not new, as they go back to some of the original research in the form of Burgess expectancy tables and the Glueck predictive tables.) The authors cite numerous studies from the 1980s that detected subgroups

within research populations having higher base rates of future violence. They state,

> These newer studies, of course, do not provide scientific evidence with which to claim validity for predictive testimony in dichotomous form ... They merely provide research support, in some cases, for predictive testimony that offers courts a sense of the relative risk of violence associated with individuals in question ... Yet this is enough to contradict the generalized assertion that all predictive testimony regarding future violence is unethical for lack of a scientific basis.

In the authors' discussion on statutes such the New York Family Court Law, they argue that risk predictions, not dichotomous predictions, are required and that detention is based on sufficient risk. Because, as they state, predictions regarding the degree of risk have been shown to be accurate, the degree of risk deemed sufficient is up to society. However, given that a threshold has been set to justify preventive detention, say at 40% probability of committing an act of violence, the decision remains dichotomous; as Gottfredson (1987) points out, "Most selection applications of prediction devices use some cutting score that essentially reduces the predictor scale to a dichotomy" (p. 30).

2.8.1 The Bail Reform Act of 1984

Repealing the Bail Reform Act of 1966, Congress passed the Bail Reform Act of 1984 (H.R. 5865, 1984). The new version, among other things, authorized judicial officers to determine pretrial detention (§ 3141(a)), required a detention hearing held in any case involving a violent crime (§ 3142(f)(1)(A)), and required detention of any person awaiting sentencing unless the judicial officer determined, with clear and convincing evidence, the person was *not* likely to flee or did not pose a threat to the community or any persons (§ 3142(f)). Shortly after, the constitutionality of the Bail Reform Act of 1984 (specifically the clauses mentioned) was challenged and reached the United Stated Supreme Court (*United States v. Salerno*, 1987). The Supreme Court ruled that detention under the Act was regulatory

and not penal. In his dissent, Justice Thurgood Marshall stated, "The majority's technique for infringing [the right to be free from punishment before conviction] is simple: merely redefine any measure which is claimed to be punishment as 'regulation,' and, magically, the Constitution no longer prohibits its imposition" (p. 765).

In Louisiana, an individual is committed to a psychiatric hospital after being found guilty by reason of insanity unless that individual can prove that he or she is not dangerous. If committed, that individual can be released if no longer posing a threat to others or himself/herself and upon hospital recommendation, but not before the court holds a hearing to determine the individual's dangerousness. If the court finds the individual is dangerous, he/she is returned to the psychiatric hospital regardless of whether the individual is mentally ill. This process was challenged in *Foucha v. Louisiana* (1992), and the United States Supreme Court ruled that the defendant, Terry Foucha, be released. In some ways, the ruling that Foucha could not be held strictly based on perceived dangerousness seemed to contradict the ruling of *United States v. Salerno* (1987), but as Justice Byron White noted in the majority opinion, "*Salerno*, unlike this case, involved pretrial detention" (p. 81). Corrado (1996) opines that the ruling implied detention based on perceived dangerousness was not permissible; however, he also notes,

> [A]s Justice [Sandra Day] O'Connor made clear in her concurring opinion, the states may eliminate the insanity defense altogether (*Foucha v. Louisiana*, pp. 88–89, O'Conner, J., concurring), which means that people innocent of crimes because of insanity would be convicted and imprisoned nevertheless—a good functional substitute for detention on grounds of dangerousness (see Robinson, 1993).

2.9 Combining Prediction Methods

Sawyer (1966) makes the distinction between clinical and statistical (mechanical) prediction and clinical and statistical measurement; defining the former as how the data are

combined and the latter as how the data are *collected*. In combination, he notes eight prediction methods: pure clinical (collected and combined clinically), pure statistical (collected and combined statistically), trait ratings (collected clinically, combined statistically), profile interpretation (collected statistically, combined mechanically), clinical composite (collected both clinically and statistically, combined clinically), mechanical composite (collected both clinically and statistically, combined statistically), and two forms of syntheses. The first he called clinical synthesis, defined as "[t]ak[ing] a prediction, produced by *mechanical* [original emphasis] combination, and treat it as a datum to be combined *clinically* [original emphasis] with the other data"; the second, mechanical synthesis, he defined as "[t]ak[ing] a prediction, produced by *clinical* [original emphasis] combination, and treat it as a datum to be combined *mechanically* [original emphasis] with the other data." (p. 184). Sawyer examined 45 studies and concluded, among other things, that statistical combinations outperform clinical combinations for every type of data collection, and within each type of data combination the clinical mode of data collection is inferior to the other types of data collection.

The debate on clinical versus actuarial prediction performance is ongoing, but Monahan et al. (2001) feel it should have been declared over long ago: "More research demonstrating that the outcome of unstructured clinical assessments ... seem[] to be overkill: That horse [i]s already dead" (p. 7; for arguments supporting clinical predictions of violence in the second generation, see Lidz, Mulvey, & Gardner, 1993; Litwack, 2001). Although the general conclusion among researchers is that clinical prediction is inferior to statistical predictions, Gardner, Lidz, Mulvey, and Shaw (1996b), in their "progress report" on actuarial prediction of violence, note that actuarial methods are seldomly used in practice. Some suggested clinical prediction should be combined with actuarial methods (e.g., in general, see Kleinmuntz, 1990; specifically for violence prediction, see Buchanan, 1999). G. T. Harris et al. (1993) state,

> If adjustments [to actuarial methods] are made conservatively and *only* [original emphasis] when a clinician believes, on good evidence, that a factor is related to

> the likelihood of violent recidivism in an individual case, predictive accuracy may be improved. (p. 333)

However, they have since recanted this belief (Quinsey et al., 2006; M. E. Rice, Harris, & Hilton, 2010) and the authors of the VRAG take the extreme view: "What we are advising is not the addition of actuarial methods to existing practice, but rather the replacement of existing practice with actuarial methods" (Quinsey et al., 2006, p. 197; see also Quinsey, Harris, Rice, & Cormier, 1999, p. 171).

Helmus, Hanson, and Thornton (2009) report that the recidivism norms of the Static-99 (i.e., the rates of recidivism for the risk categories) needed revision, citing declining crime rates as a reason. Sreenivasan, Weinberger, Frances, and Cusworth-Walker (2010) discuss the consequences of the changing norms and question the importance given the Static-99 by the courts. Because of the flaws that may exist with Static-99 norms, the authors pushed for clinical judgment to assist in decision making. But Abbott (2011) suggests their arguments are incorrect and without empirical support and that supplementing clinical judgment will only worsen the situation.

In his seminal work, *Clinical Versus Statistical Prediction*, Meehl (1954) suggested that a clinician, in certain circumstances, can provide valuable information that may not be included in an actuarial device. He provided the following example:

> [S]uppose that ... we are trying to predict whether a given professor will attend the movies on a given night. On the basis of the [data] and a failure to show any time-series change in the relative frequency when the occasions are ordered as to time, we arrive at a probability of .90 that he will attend the neighborhood theater, the present night being Friday. The clinician, however, knows in addition to these facts that Professor A has recently broken his leg. This single fact is sufficient to change the probability of .90 to a probability of approximately zero. (pp. 24–25)

The example provided by Meehl (1954) is often used as justification for combining statistical and clinical prediction. But Meehl later cautions clinicians to "beware of overdoing

the broken leg analogy" (1957, p. 270). Although circumstances may not be as obvious as in the example provided by Meehl, clinicians may nevertheless hold strong opinions on adjustments. In other words, misplaced intuition may again be the clinician's own undoing, despite the use of actuarial methods. In testing this idea, Hilton and Simmons (2001) found that in tribunal decisions about mentally-ill prisoners, actuarial risk assessment did not influence clinical judgment and tribunal decision making. Krauss and Sales (2001) found that clinical testimony is more influential than actuarial results on jurors' decisions. Ægisdóttir et al. (2006) found in their meta-analysis that access to statistical methods failed to improve a clinician's accuracy, and possibly worsened it. That's not to say clinicians have no value to prediction; as Hilton et al. (2006) put it, "Because some of the best indicators [for predicting violence] require clinical skill to measure, accurately appraising violence risk is likely to remain a task for the clinician, but the place for the clinical judgment is within rather than outside actuarial tools" (p. 402). Dawes (2002) argues that when a validated, superior statistical prediction measure exists, practitioners have an ethical obligation to use them (see also Dawes, 2005); he also states,

> [T]he practitioner claiming to use his or her own intuition to "improve" [a statistical prediction method] has an ethical obligation to keep track of outcomes to see if modification really does result in improvement (and must be wary of confounded judgments, self-fulfilling prophesies and other challenges to the validity of evaluating such feedback). (pp. S181–S182)

And what about the combination of multiple actuarial instruments? Seto (2005) notes that combining instruments does not increase accuracy and in some instances reduces it. Taking a different approach to violence prediction, Skeem, Manchak, Lidz, and Mulvey (2013) suggest that an accurate measure of risk assessment is to simply ask the individuals (i.e., measure risk via self-reported measures). Their results show that a self-perception measure of violence is a significant predictor of serious and any violence, even outperforming two risk assessment measures (one being the COVR). The authors' results also suggest that

self-perception adds incremental utility to the RAIs but not the other way around.

2.10 Aggregate Versus Individual Prediction

S. D. Hart, Michie, and Cooke (2007) discuss two types of errors when using risk assessment instruments (RAIs): group (i.e., nomethetic or aggregate) and individual (i.e., idiographic) error. Group error typically is expressed in the form of confidence intervals (CIs). Individual error is a bit more complicated, and, as the authors note, rarely provided. The authors suggest using individual CIs that can be calculated using a method first introduced by Wilson (1927) and given below:

$$\frac{\hat{p} + {z^2_{\alpha/2}}/{2n}}{1 + {z^2_{\alpha/2}}/{n}} \pm \frac{z_{\alpha/2}\sqrt{{\hat{p}(1-\hat{p})}/{n} + {z^2_{\alpha/2}}/{4n^2}}}{1 + {z^2_{\alpha/2}}/{n}},$$

where $\hat{p}$ is an estimate for the proportion of individuals with a given RAI score who recidivate (or, are violent, violate parole, etc.), $z_{\alpha/2}$ is the $1 - \alpha$ quantile of the standard normal distribution (i.e., $P(Z \leq z_{\alpha/2}) = \alpha/2$), and n is the number of individuals with a given RAI score. Agresti and Coull (1998) note that the coverage of Wilson's method for estimating a confidence interval is closer to the $1 - \alpha$ value than other methods such as Wald's method ($\hat{p} \pm z_{\alpha/2}\sqrt{{\hat{p}(1-\hat{p})}/{n}}$). As S. D. Hart et al. (2007) note, Wilson's method requires few assumptions, such as the reasonable assumption that RAI scores follow the binomial distribution; is easy to calculate; and does not require the raw data, only the individually-assigned probabilistic value. S. D. Hart et al. (2007) examined individual CIs for the Static-99 and VRAG and found that the individual intervals were very large, especially compared to the group CIs, for these RAIs. For instance, the average width of a 95% confidence interval for the different Static-99 categories was about 13 percentage points and for the VRAG, 20 percentage points. For the individual 95% CIs, the average widths were 86 and 85 percentage points, respectively, for the Static-99 and VRAG. As an example, an individual who scores

an 8 on the VRAG is categorized in a group where 76% of the individuals are recidivists. The 95% CI for this group estimate is (.58, .88); for the individual estimate using Wilson's method, it is (.12, .99). In other words, for individuals who score an 8 on the VRAG one can be confident that most (about 3 of every 4) will recidivate; however, for any given individual who scores an 8, one cannot be as confident. It is worth noting that when $n = 1$, the width of a 95% CI using Wilson's method depends strictly on the point estimate (i.e., estimated probabilities); the width is minimal at 0 and 1 (.79) and maximal at .5 (.89). Even a 50% confidence interval will have a width larger than .5 for more than half the possible point estimates. This is not a criticism of Wilson's method but rather a reminder of the inherent difficulties of individualized prediction. S. D. Hart et al. (2007) caution against the use of RAIs in situations where decision errors are high; in addition, they state that "[l]ow predictive accuracy not only makes reliance on [actuarial]RAIs ethically problematic, it also means that they may not meet legal standards for the admissibility of expert or scientific evidence" (p. s64).

S. D. Hart et al. (2007) sparked controversy with their article; in their response G. T. Harris, Rice, and Quinsey (2008), argue that actuarial methods, although infallible, are the best available:

> An actuarial tool ... is simply an efficient distillation of relevant empirical evidence. Actuarials do not afford certainty, but ... are more valid than any other method. The *undeniable* [emphasis added] superiority of actuarials means that their use can optimise the balance between public safety and offenders' civil liberties. Hart et al's advice to eschew risk-related decisions means less accurate decisions that cumulate in avoidable harm to victims, unnecessary restriction of offenders, or both. (p. 154)

G. T. Harris and Rice (2007) also commented on the S. D. Hart et al. (2007) article, stating that their article "does not sensibly indict actuarial tools or any other empirically based decision policies using data from groups" and that

> the psychometric data about the (VRAG) (e.g., Harris et al., 1993; Harris et

> al., 2003) have indicated that its standard error of measurement is such that, with 95% confidence, an individual true score differs from an obtained score by less than one category. (p. 1648)

Mossman and Sellke (2007) question, "[I]f all one knows about an individual is his membership of a risk group, what can 'individual risk' mean?" The authors suggest that "95% CIs for 'individual risk' pile nonsense on top of meaninglessness" (p. 561). Skeem and Monahan (2011), expressing their opinion on the matter, state that "group data theoretically can be, and in many areas empirically are, highly informative when making decisions about individual cases" (p. 40). They support their argument by analogizing human behavioral risk with insurance sales and rainy days.

In their article discussing the use of the OGRS used for assessing risk of harm and reoffense, Copas and Marshall (1998) note,

> The statement accompanying the OGRS is careful to point out that the score is not a prediction about an individual, but an estimate of what rate of conviction might be expected of a group of offenders who match that individual on the set of covariates used by the score. (p. 170)

Grove and Meehl (1996) also discuss the issue with aggregate prediction:

> There is a real problem, not a fallacious objection, about uniqueness versus aggregates in defining what the statisticians call the reference class for computing a particular probability in coming to a decision about an individual case (p. 306)

They argue that the probability of an outcome attached to an individual depends on which reference class is used. This is a notion echoed by Janus and Meehl (1997). They suggest, as do others (e.g., Monahan, 1973; Meehl, 1954) that individual probabilities are, in reality, statements about the group to which a person belongs. For example, stating that an individual has a .75 probability of recidivism is equivalent to stating that the individual belongs to a (reference) group where 75% of the individuals recidivate.

Berlin, Galbreath, Geary, and McGlone (2003) express caution as well, stating, "It is critical to note that individuals within any such 'high risk' group are not all at equal risk" (p. 378). G. Harris (2003) criticizes the work of Berlin et al. (2003), calling it "essentially a layperson's commentary on the recent research in the field of risk assessment" (p. 389). Specifically addressing the idea Berlin et al. presented in the above quote, G. Harris (2003) states,

> [V]irtually all decisions require clinicians to treat patients or clients as members of groups. Diagnosis and prognosis inevitably demand probabalistic statements about whether an individual falls within a particular reference class (even if clincians address the probabilities only implicitly). Attempting to treat a patient or client as if he were unique in any real sense (i.e., not a member of any group) would require clinicians to ignore all prior scientific research. (p. 391)

2.10.1 Prediction Intervals

Cooke and Michie (2010) suggest using prediction intervals, rather than confidence intervals, when providing an interval estimate for an individual's risk. Prediction intervals are used to convey uncertainty in prediction of a new observation; they depend on the variability of both the estimated model and new observation. Prediction intervals are wider than comparative confidence intervals and depend less on the size of the sample. Consider the simple linear regression model, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. The confidence interval for true expected value of y given $X = x_*$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_* \pm t_{\alpha/2,n-2} \left(\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x}^2)}{\mathrm{SS}_x} \right)} \right),$$

where $\hat{\sigma}^2$ is the estimated error variance, SS_x is the sum of squares of the mean deviations for x, and $t_{\alpha/2,n-2}$ the critical value of a T-distribution with $n-2$ degrees of freedom such

that $P(T \leq t_{\alpha/2,n-2}) = \alpha/2$. For a new observation, the prediction interval is

$$\hat{\beta}_0 + \hat{\beta}_1 x_* \pm t_{\alpha/2,n-2} \left(\sqrt{\hat{\sigma}^2 \left(\frac{n+1}{n} + \frac{(x_* - \bar{x}^2)}{\mathrm{SS}_x} \right)} \right).$$

Cooke and Michie (2010) estimated prediction intervals for the probability of reconviction estimated by a logistic function and based on PCL-R scores. The 95% prediction intervals for different PCL-R scores did not deviate much from the uninformative $(0, 1)$ interval. Cooke and Michie (2010) conclude, "[O]n the basis of empirical findings, statistical theory, and logic it is clear that predictions of future offending cannot be achieved, with any degree of confidence, in the individual case" (p. 272).

Attempting to address many of the criticisms received for the S. D. Hart et al. (2007) article, S. D. Hart and Cooke (2013) built a logistic regression model; the individual risk estimates from the model ranged from .03 to .68 with confidence interval widths ranging from .14 to .64 indicating, according to the authors, that "the individual risk estimates made using the [actuarial]RAI had margins of error that were very large, indicating that it was virtually impossible to identify subjects with distinct or non-overlapping probabilities of failure" (p. 94). S. D. Hart and Cooke (2013) also question how RAIs can be considered legally admissible under the *Daubert* standard "when the margins of error for individual risk estimates made using the tests are large, unknown, or incalculable" (p. 97).

Hanson and Howard (2010) disagree with S. D. Hart et al. (2007) and argue that an individual "confidence interval of $[0, 1]$ is a consequence of having only two possible outcomes" (p. 279; i.e., due to a dichotomous outcome variable). Because of this, they state that "individual confidence intervals are uninformative about the accuracy of the risk assessment procedure when the outcome is dichotomous" (p. 280).

2.10.2 Credible Intervals

Scurich and John (2012) attack the controversy altogether for being centered around

the "frequentist" approach using confidence intervals; they state, "[i]t is simply unintelligible to employ frequentist confidence intervals to describe the precision of actuarial risk estimates" (p. 243). Similar to Hanson and Howard (2010), they also dismiss the idea of using a prediction interval for a binary outcome, such as whether an individual will recidivate, because "intervals around discrete quantities make no sense" (p. 239). Instead, the authors advocate the use of Bayesian credible intervals. Using the same VRAG data provided in S. D. Hart et al. (2007), Scurich and John (2012) provide a numerical example using four different beta distributions for the Bayesian prior of the parameter, p, representing the probability of recidivism. The probability density function for the beta distribution is provided in Equation (2.1):

$$f(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \tag{2.1}$$

where $\Gamma(\cdot)$ is the gamma function, $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\mathrm{d}x$. The choice for the parameters α and β for the four priors used by Scurich and John were representative of four different hypothetical people: an naïve person ($\alpha = \beta = 1$; i.e., the standard uniform distribution, an uninformative prior) and three mental health professionals working in (a) an outpatient setting ($\alpha = 1.5$, $\beta = 3.5$), (b) a facility with involuntarily- and voluntary-committed patients ($\alpha = 3$, $\beta = 7$), and (c) a maximum-security ward ($\alpha = 15$, $\beta = 35$). The expected value of the beta distribution given in Equation (2.1) is $\alpha/(\alpha+\beta)$; thus, all three informative priors have the same mean, .30, representing the approximate base rate in the sample. The four credible intervals for an individual scoring an 8 on the VRAG is, for the naïve person, $(.58, .88)$ and for the three mental health professionals, (a) $(.53, 83)$, (b) $(.49, .78)$, and (c) $(.36, .58)$. It is clear from the four intervals that the prior distribution plays an important role, which the authors note, "deserves some attention within the context of violence risk assessment"; unfortunately, this was beyond the scope of their article. Given a proper way of constructing a prior distribution, a credible interval offers the distinct advantage of attaching

a probability to the parameter p, whereas a confidence interval only provides the confidence one has that the interval will, in the long run, contain p.

2.11 Predictors of Recidivism

The type of predictor variables used in risk assessment measures can vary but commonly include the perpetrator's age, criminal history (e.g., number of previous arrests), victim (particularly used with sex offenders), substance abuse, and personality (e.g., the Psychopathy Checklist). Predictor variables are often classified into two categories, dynamic and static, described in the next section.

2.11.1 Dynamic Versus Static Predictors

Predictor variables that change over time are referred to as *dynamic predictors*. As Dr. Howard Barbaree states in his comments in Grisso, Malamuth, Barbaree, Quinsey, and Knight (2003), "Static factors are a between-subjects kind of variable and dynamic factors are a within-subject variable" (p. 240). Examples of dynamic predictors include current or recent substance abuse, recent death in family, and current treatment. In contrast, *static predictors* are ones that do not change over time, or change in one direction. Examples of static variables include age, many of the variables related to criminal history (e.g., number of previous arrests) but not all (e.g., months since last arrest), many of the variables related to the victim (e.g., victim's age and gender at time of offense), and family history variables (e.g., father's drug use). Some variables do not fit neatly into the two categories; for example, many personality traits are considered stable but not until after a certain age, and for some traits this occurs quite late in one's life (B. W. Roberts, Walton, & Viechtbauer, 2006).

Gottfredson and Moriarty (2006) opine that static variables are more common because dynamic variables "may be more problematic from a methodological viewpoint because such variables may be more difficult to measure" (p. 192). Quinsey et al. (2006) argue that

dynamic predictors pose numerous problems. For instance, they suggest that "predictors that change may lose their predictive ability with remeasurement" (p. 45) and state that it is not possible to compare the accuracy of static predictors with dynamic predictors "unless one uses dynamic predictors as measured at a particular point of time" (p. 45). G. T. Harris and Rice (2003) state,

> Predictive accuracy achieved by already-identified static, historical variables places a limit on the amount of remaining variance available for assessments (one-time or change scores) of such putatively dynamic variables as changes in mood, insight, or procriminal attitudes. We conclude that in predicting long-term violent recidivism, there is very little outcome variance left over for "dynamic" variables. ... As yet, there have been no demonstrations that change scores add anything to initial scores in predicting recidivism among sex offenders. The predictive accuracy achieved under optimal conditions by a comprehensive set of such static, historical predictors ... imply that such dynamic attitudinal and intrapsychic variables (assessed one or more times before release) cannot make *incremental* [original emphasis] contribution to the prediction of recidivism. (pp. 204–205)

Under the assumption that psychopathic personality is static, the Violence Risk Appraisal Guide is an example of a risk assessment tool that is based on only static variables.

But just because a measure uses static variables does not mean it will necessary produce similar results from different evaluators. Consider the Static-99, a tool that has been shown to possess interrater reliability (e.g., de Vogel, de Ruiter, van Beek, & Mead, 2004) and contains only static predictors (see Appendix B). Boccaccini et al. (2012) looked at Static-99 evaluations in Texas and New Jersey and found that rater agreement (as measured by intraclass correlations) were strong (.79 in Texas and .88 in New Jersey), but that evaluation pairs were identical only about half the time (55.0% in Texas and 54.1% in New Jersey). When looking at the scoring at the item-level, evaluation pairs were in agreement across all items less than half the time (49.7% in Texas and 43.0% in New Jersey). The disagreement

in total scores was generally within one point, but over ten percent of the evaluation pairs were more than two points (12.5% in Texas and 12.6% in New Jersey). The authors suggest the use of confidence intervals around raters' total scores. C. S. Miller, Kimonis, Otto, Kline, and Wasserman (2012) found similar results with the PCL-R, MnSOST-R, and Static-99.

In *United States v. Abregana* (2008), convicted sex offender Jay Abregana received evaluations from three experts, Dr. Dennis M. Doren, Dr. Luis Rosell, and Dr. Howard E. Barbaree. Abregana received different scores on both the RRASOR and Static-99 (they could not agree on the shared item "prior sex offenses"). It should be noted that Drs. Rosell and Barbaree—both witnesses for the defendant—did agree, whereas Dr. Doren, a witness for the United States, had a higher overall score. In addition, the three agreed on only twelve of the sixteen MnSOST items and all three overall scores differed.

An actuarial device that provides a probabilistic measure based on static variables implies that the individual's risk for violence remains static. As Slobogin (2006) points out, the score on an actuarial instrument such as the VRAG, "does not take into account whether [an individual] is undergoing treatment, is about to get married, has recently lost functioning in one or more limbs, or has found religion" (pp. 23–24). Bechman (2001) says, "There is little dispute that a consideration of dynamic factors can be important in accurately assessing risk" (p. 28). Some research has supported this; Gendreau, Little, and Goggin (1996) found in their meta-analytic study that many dynamic predictors are as good and often better than static predictors.

Instruments that use dynamic predictors assume that risk is liable to change. This point is made by Douglas and Skeem (2005) where they distinguish between what they call *risk status* and *risk state*; as they explain, "Static risk factors describe an individual's risk status, whereas a combination of static and dynamic factors describes an individual's risk state" (p. 350; see also Skeem & Mulvey, 2002). Some research (Skeem et al., 2006, e.g.,) has shown that changes in some psychiatric symptoms (e.g., anger) may lead to an increased risk of violence. Gottfredson (1987) plainly states, "If a variable can be measured reliably

and if it is predictive, then of course it should be used—absent legal or ethical challenge" (p. 193).

2.11.2 Gender and Race

One of the factors used in Tibbitts (1931) was "Nationality" and it was noted that "Irish," "Austrians," and "Negroes" are more likely to violate parole. The use of race or ethnicity in the prediction of dangerousness, whether directly or indirectly, is a controversial topic. Such a factor would likely never explicitly be included in today's actuarial models (e.g., Monahan et al., 2001, in the development of the COVR state that they removed race from their model "on ethical and legal grounds" [p. 119]).

Using race as a predictor involves the controversial issue of racial profiling. In 1996, Victor Saldano was sentenced to death for capital murder; an expert witness in the case was a psychologist named Dr. Walter Quijano, who testified on behalf of the state. Dr. Quijano assessed Saldano's future dangerousness using numerous risk factors, including race, and Saldano was ultimately sentenced to death (*Saldano v. Roach*, 2004; see also Monahan, 2006). After certiorari was granted, the Attorney General of Texas, before the United States Supreme Court, confessed error in introducing race as a factor; however, the en banc session of the Texas Court of Criminal Appeals concluded that the confession of error was "contrary to [Texas's] procedural law for presenting a claim on appeal" (*Saldano v. State*, 2002, p. 891). The Court did not collectively opine on the use of ethnicity or race in determining risk of dangerousness, but several dissenting members of the Court did. For example, Judge Tom Price stated,

> Dr. Quijano's testimony during the punishment phase of the appellant's trial drew a correlation between the appellant's race and incarceration rates. I would hold that the admission of this evidence was fundamental error, which should be reviewed even in the absence of a trial objection. (p. 892)

Judge Cheryl Johnson stated, "I do not think that race or ethnicity should ever be a consideration, in any degree, in the assessment of punishment" (p. 893). The legislators of the state of Texas later addressed the issue of race: "evidence may not be offered by the state to establish that the race or ethnicity of the defendant makes it likely that the defendant will engage in future criminal conduct" (Texas Code of Criminal Procedure, 2014).

In another Texas capital murder case, *Buck v. Thaler* (2011), Dr. Quijano again testified in court, this time on behalf of the defendant, Duane Buck, an African-American. Dr. Quijano testified that the defendant did not pose a threat to society; however, when the defense questioned Dr. Quijano about statistical factors used, he cited race as one and suggested that Mr. Buck's race is a factor related to increased risk of dangerous behavior.

Although race may not be explicitly provided in predictive measures, it is often implicitly included; as Berk (2009) states, "In locales with substantial residential segregation, knowing the zipcode is virtually the same as knowing an individual's race" (p. 232). Professor Bernard Harcourt says,

> When you live in a world in which juveniles are much more likely to be stopped—or, if stopped, be arrested, or, if arrested, be adjudicated—if they are black, then all of the indicators associated with prior criminal history are going to be serving effectively as a proxy for race. ... [It] inscribes the racial discrimination you have today into the future. (Labi, 2011, para. 9)

Berk (2009) examined the role of race in parole failures; using sophisticated decision tree analyses (specifically random forests; see Chapter 5), he found that race plays an important role in the model's predictive accuracy. Berk (2012) later argues, "By not including age, gender, and race as forecasting predictors, one may be sparing some young African-American men substantial time in prison, but at the cost of the deaths of other young African-American men" (p. 8). Heilbrun, Douglas, and Yasuhara (2009) state that research is lacking in terms of risk assessment measures performance across different ethnicities; lacking a predictor variable for race may render risk assessment measures that are developed

on racially-homogeneous populations useless. For instance, Långström (2004) found that the Static-99 and RRASOR could not distinguish recidivists and non-recidivists among a Swedish subpopulation of African-Asian descent.

Many risk assessment tools are constructed using predominantly male samples (e.g., the construction sample for the VRAG was all male) and some research has shown that they may not be appropriate for use in female populations (e.g., de Vogel & de Ruiter, 2005; G. T. Harris et al., 2002). Most crime statistics support the notion that males are far more likely to commit violent acts than women (e.g., Maccoby & Jacklin, 1974).

It is clear that more research is needed before making conclusions, and until conclusions suggest that risk assessment measures generalize across race and gender, evaluators should take extreme caution when assessing select populations.

2.12 A Third Generation?

Several authors have suggested a third generation of violence risk is in effect (e.g., Bonta, 1996), consisting of risk assessment measures that are "theoretically based, inclusion of dynamic items, and concerned with measuring changes in risk" (Campbell, French, & Gendreau, 2007, p. ii).

2.12.1 Structured Professional Judgment

Skeem and Monahan (2011) suggest that risk assessment methods "now exist[] on a *continuum of rule-based structure* [original emphasis']" (p. 39) as opposed to the simple clinical/actuarial dichotomy. As they suggest, the two extremes of this continuum are completely unstructured (i.e., clinical) and completely structured (i.e., actuarial) assessments. Somewhere in the middle of the continuum exists what some call *structured professional judgment* (SPJ; e.g., see Kropp & Hart, 2000) where "decision-making is assisted by guidelines that have been developed to reflect the 'state of the discipline' with respect to scientific

knowledge and professional practice" (S. D. Hart, 2009, p. 150).

A completely structured actuarial assessment tool, such as the Violence Risk Appraisal Guide (VRAG), advises against any sort of modification based on clinical judgment because, as M. E. Rice et al. (2010) state, "[N]o evidence supports contentions that any alterations (based on clinical judgment) of actuarial scores ... results in more accurate decisions compared to actuarially derived scores alone" (p. 99).

S. D. Hart (2009) notes the difference between the Risk for Sexual Violence Protocol (RSVP; an SPJ assessment tool) with the Static-99 (an actuarial assessment tool): "Rather than offer a prediction of what will happen, the evaluator [using the RSVP] speculates systematically about what the offender might or could do in the future and how to prevent it. ... [The Static-99] is focused on prediction, not prevention" (pp. 163–164). As Heilbrun et al. (2009) describe it,

> Rather than reaching a numeric conclusion, SPJ evaluators are asked to rate each case as low, moderate, or high risk by assuming (1) the more risk factors present that are *individually relevant to a person's violent behavior* [original emphasis], the higher the risk; and (2) the greater the degree of intervention required to stem the risk of violence, the higher the risk. ... The "judgment" part of SPJ requires evaluators to decide for whom it is relevant (and to what extent) and for whom it is not. (p. 337)

As Doyle and Logan (2012) say, actuarial measures "are derived from a prediction rather than a prevention paradigm, so that conceptually they are designed to determine 'if' a person is a risk rather than 'why' (p. 411).

Along with the RSVP, the Spousal Assault Risk Assessment Guide (SARA), the Historical Clinical Risk–20 (HCR-20), the Sexual Violence Risk–20 (SVR-20), the Short-Term Assessment of Risk and Treatability (START), the Early Assessment Risk List for boys (EARL-20B), the Early Assessment Risk List for girls (EARL-21G), and the Structured Assessment of Violence Risk in Youth (SAVRY) are all considered SPJ assessment instruments

(Guy, 2008). Many of these instruments have been shown to predict violent recidivism with moderate accuracy (e.g., HCR–20, de Vogel & de Ruiter, 2006; de Vogel & de Ruiter, 2005; and Douglas, Ogloff, & Hart, 2003; SAVRY, Catchpole & Gretton, 2003; EARL-20B, Enebrink, Långström, & Gumpert, 2006), sometimes outperforming actuarial instruments (e.g., de Vogel et al., 2004, found the SVR–20 was more accurate than the Static-99 in predicting sexual recidivism; Douglas, Yeomans, & Boer, 2005, found the HCR–20 was as good and often better than many actuarial methods, including the VRAG, in predicting violence).

2.13 Violence Risk Communication

Following the ruling in *Tarasoff v. Regents of the University of California* (1976), Monahan (1993) advised,

> Four tasks form the basis of any professionally adequate risk assessment: The clinician must be educated about what information to gather regarding risk, must gather it, must use this information to estimate risk, and, if the clinician is not the ultimate decision maker, must communicate the information and estimate to those who are responsible for making clinical decisions. (p. 242)

With respect to communication, Monahan adds, "Only by making the information salient can one be assured that the decision maker has had the option to make use of it" (p. 245).

Risk communication "involv[es] the provision of information from an assessor to a decision-maker regarding the risk of a specified event's occurrence" (Heilbrun, Dvoskin, Hart, & McNiel, 1999, p. 91); it plays an important role in legal decision making. Monahan and Steadman (1996) opine, "Understanding how best to communicate assessments of risk is as important to mental health law as improving the validity of those assessments themselves" (p. 938) and suggest that risk communication closely follow the paradigm found in meteorology (e.g., no threat, violence watch, violence warning).

Heilbrun, Dvoskin, et al. (1999) argue for more research in violence risk communication and provide guidelines for communicating risk of violence in the mental health community. In a survey involving fifty-five participants who had received "Basic Forensic Evaluator Training through the Institute of Law, Psychiatry and Public Policy at the University of Virginia" only one indicated that numerical figures, such as probabilities, were used when communicating risk; about half of the participants stated they do not use numerical figures is because "the state of the research literature doesn't justify using specific numbers" (Heilbrun, Philipson, Berman, & Warren, 1999, p. 399). In a separate survey, Heilbrun, Philipson, et al. (1999) report that, among the fifty-seven participants (mental health clinicians who were participants in a continuing education workshop on risk assessment), only four indicated they used numerical probabilities to communicate risk. Hanson, Lloyd, Helmus, and Thornton (2012) suggest the use of non-arbitrary metrics for communicating risk: "[I]t is the responsibility of the professional community to establish plausible definitions for risk communication. Given that nominal categories [i.e., high-, moderate-, and low-risk] are unlikely to go away, we should work towards giving nominal risk categories explicit, non-arbitrary meanings" (p. 19). If using percentile rankings, the authors recommend also including recidivism base rates.

Heilbrun, O'Neill, Strohman, Bowman, and Philipson (2000) presented eight vignettes describing hypothetical patients differing in three aspects: risk level (low risk vs. high risk), risk factors (static risk factors [e.g., age] vs. dynamic risk factors [e.g., substance abuse]), and risk model (prediction with legal decision being civil commitment vs. management with legal decision be parole). The vignettes concluded with six different forms of risk communication and seventy-one psychologists/sociologists and psychiatrists were asked to rate (on a five-point Likert scale) how valuable they found the risk communication to be. Risk level was found to be significantly different among all respondents; in particular, risk communication identifying risk factors and specifying possible interventions was rated significantly more valuable among high-risk patients. Among low-risk patients, ratings of value differed

significantly between psychiatrists and psychologists/sociologists when the risk communication described relevant clinical characteristics (psychiatrists' ratings were higher; there were no other significant differences between the two groups of respondents). The authors conclude from the results of their study that "violence risk communication is valued by experts according to both the risk level of the individual being assessed and the nature of the risk factors that are present ... particularly true in the high-risk cases" (p. 145).

Interested in laypersons' and clinicians' perceptions of a person's risk of violence, Slovic and Monahan (1995) asked participants to judge the probability that a person would be violent, whether that person should be labeled as "dangerous," and the necessary course of action. Their results showed that the higher the perceived risk, the more likely participants (both laypersons and clinicians) advocated hospitalization (coerced, if necessary). The authors note that "the concept 'probability of harm' was not represented in [the participants' (both laypersons and clinicians)] minds in a consistent, quantitative way" (pp. 61–62). But Slovic, Monahan, and MacGregor (2000) caution that one should "temper deriving strong conclusions about risk communication" (p. 273) from these results.

Pescosolido, Monahan, Link, Stueve, and Kikuzawa (1999) surveyed U.S. residents' opinions regarding mental health in the country using vignettes describing an individual who was either schizophrenic, severely depressed, alcohol-dependent, drug-dependent, or troubled/distressed (used as the control). The authors discovered, among other things, that respondents viewed most individuals as being somewhat or very likely to be violent toward others (lowest: 17% among the troubled group; highest: 87% among the drug dependent group). The authors also found that when the person was considered dangerous to others that coerced treatment was deemed necessary (lowest: 83% among the troubled group; 96% among the drug dependent group).

Slovic et al. (2000) conducted a three-part study: Study 1 asked experienced clinicians to estimate violence risk after reviewing hospital discharge summaries; Study 2 instructed clinicians in making probabilistic judgments and compared their violence risk estimates to the

(uninstructed) estimates from Study 1; Study 3 measured the effects of risk communication in clinicians' decision making. In Studies 1 and 2, the authors also examined the effect of communicating risk with probabilities versus per-one-hundred frequencies stated using large-response scales or small-response scales, creating four distinct groups (in Study 1, there was a fifth group that were presented small-response scale frequencies per one-thousand). In Study 1 they found that when likelihood judgments were stated as frequencies they were significantly lower than when stated as probabilities; however, within the different likelihood categories (e.g., ".01", "1 in 100", "10 in 1000"), respondents who stated their likelihoods in frequencies were more likely to label the patient as medium- or high-risk. When provided with a tutorial (Study 2) the differences in mean likelihood estimates were greater when presented in terms of probabilities, but only between the large-response scales groups. The mean likelihood judgments in Study 2 were uniformly higher across the four groups, but the tutorials "did not substantially reduce or eliminate scale [i.e., large vs. small response scales] and format [i.e., probability vs. frequency] effects" (p. 283). Again, those in the frequency groups labeled more patients as medium- or high-risk across nearly every likelihood level. In Study 3, there were seven questionnaires differing only in the assessment of a hypothetical patients (see Fig. 5, p. 297 in Slovic et al., 2000), and participants were asked to make several judgments (e.g., the risk of harming one's self or others). Once again, their findings showed that when presented risk of violence in terms of frequencies, the patient was more likely to be labeled as medium- or high-risk. Slovic et al. (2000) conclude, "Clearly, it makes no clinical or policy sense to keep twice as many people in the hospital when their risk of violence is characterized as '20 out of 100' than when it is characterized as '20%' " (p. 292).

In a similar study, Monahan et al. (2002) compared risk communication using probability versus frequency formats and pallid versus vivid outcome depictions. Participants—all whom were members of the American Psychological Association with an interest in clinical or forensic psychology—were told that a hypothetical patient had just been evaluated for discharge. Participants in the pallid outcome were told that "recently, a stranger in the

community sustained fatal injuries from another patient who was discharged from the same facility" (p. 122); those in the vivid outcome were told that "recently, another patient who was discharged from the same facility killed a stranger in the community by smashing her skull with a baseball bat, resulting in her instant death" (p. 122). Each participant was then presented with a risk estimate in probability or frequency format and asked if the patient (a) should be discharged, (b) not be discharged, or (c) if a second opinion was needed. Similar to the results of Slovic et al. (2000), the authors found that those participants in the frequency format were more conservative in their decision to discharge a patient. Additionally, the authors found that those in the pallid outcome/probability format group were the least conservative and those in the vivid outcome/frequency format were the most conservative. However, their results were only significant for participants who worked in forensic facilities. In noting relevant differences between the two groups, the authors stated that the participants who work in forensic settings "provide consultation more often to courts, and ... provide formal assessments of dangerousness more often as part of their practice" (p. 126). In other words, these participants are the ones more likely to make such a meaningful decision.

Monahan and Silver (2003) conducted a survey among judges using the five risk categories from the COVR (see Table 1.3 in Chapter 1). The judges were presented with five risk categories, either in frequency or probabilistic format, and asked to decide on the *lowest* likelihood of violence (according the COVR estimate) that would meet their definition of someone being dangerous to others or himself, requiring short-term civil commitment. Results from twenty-six judges (thirteen in each format group) reveal that, as shown by Slovic et al. (2000) and Monahan et al. (2002), frequency formatting led to more conservative decision making (i.e., smaller likelihood of violence needed to civilly commit). Three judges selected the very low-risk group (all in the frequency format) and no judge selected the very high-risk group. The mean group levels selected were 2.4 (for the frequency format), 2.9 (for the probabilistic format), and 2.7 (across both formats). The authors summarize their

findings as follows:

> Risk Class 3—a 0.26 likelihood of committing a violent act—best reflects the judges' decision threshold for short-term civil commitment as "dangerous to others." Put otherwise, while fully half the judges (13 of 26) were of the view that people with a serious mental disorder whose risk was assessed by the MacArthur procedures as being in Risk Classes 1 or 2 qualified for commitment, it was not until Risk Class 3 was reached that a majority of the judges (19 of 26) would commit.

In connection with the aggregate-versus-individual debate previously discussed, Scurich, Monahan, and John (2012) examined *when* people apply aggregate-level data to individuals. The authors collected data from jury-eligible mock jurors in the United States; each participant read COVR case-summary vignettes that included an estimated risk in the form of frequencies (three groups: low risk: eight out of 100; medium [average] risk: 26 out of 100; or [very] high risk: 76 out of 100; see Table 1.3 in Chapter 1) and possibly provided additional risk factors (three "packing" groups: six, three, or no risk factors given) that are among the COVR variables (e.g., father's drug use, recent violence). Participants were asked for the likelihood that they would commit the hypothetical patient and how likely they felt the patient was one of the X number of people who were violent (where $X = 8$, 26, or 76), as measured using seven-point Likert scales. In addition, the authors measured participants' numeracy using the Subjective Numeracy Scale (SNS; Fagerlin et al., 2007). An individual said to be an *innumerate* is defined as being "marked by an ignorance of mathematics and the scientific approach" (Innumerate, n.d.). Their results suggest that no unpacking (i.e., no risk factors presented) does not lead to significant differences in commitment likelihood across different levels of risk, but unpacking (i.e., three or six risk factors presented) does. In addition, for high-risk patients, unpacking led to an increased likelihood to commit, whereas in the low-risk patients the likelihood decreased. Among high-risk patients, unpacking did not significantly affect participants' perceived likelihood that the patient was one of the X number of people who were violent; however, in the low-risk group, unpacking reduced this

likelihood. Their results also suggest that numerates were not affected by the unpacking scheme.

In one of their most important works, Tversky and Kahneman (1981) show that the framing of a question, whether in terms of losses or gains, lead to predictable changes in preferences. Their 1981 article appeared in *Science* and showed that participants' choices differed among two mathematically-equivalent, forced-choice questions posed in terms of either lives saved or lives lost. Their "Asian disease problem" proposed two options, a sure thing or a gamble; when the question is posed in terms of lives saved, most participants choose the sure-thing; when posed in terms of lives lost, most choose the gamble. Scurich and John (2011) looked to see if the *framing effect* exists in the context of civil commitment decisions. The authors suggest that "[b]etter-informed decisions depend not only on the substance of the [risk] assessment, but also on how that substance is communicated to the decision maker" (p. 83). In their study, undergraduate mock-judge participants were given six COVR case-summary vignettes and presented with a forced-choice question asking if they would commit or release the hypothetical patient and indicate how likely—on a six-point Likert scale—they were to do so. The case summaries provided estimated risk probabilities (three groups: [very] low risk: .01; moderate [average] risk: .26; or [very] high risk: .76; see Table 1.3 in Chapter 1) and their corresponding confidence intervals framed in terms of either "probability of violence occurring" (e.g., .26) or "probability of violence not occurring" (e.g., .74). When framed as a probability of violence occurring, significantly more participants had patients committed and were significantly more likely to commit patients (given high and moderate probabilities of violence; there were no significant differences for the low-risk probability).

2.14 Determining the Accuracy of Predictions

Measuring accuracy has long been a topic of debate in violence and sexual recidivism prediction. In his article, Gottfredson (1987) discusses several measures of predictive accuracy. Some of the accuracy measures discussed are Pearson's chi-square (Pearson, 1900b), Pearson's ϕ (Pearson, 1900a), the index of predictive efficiency (Horst, 1941; Ohlin & Duncan, 1949), mean cost rating (MCR; Berkson, 1947; Duncan et al., 1953), and relative improvement over chance (RIOC; Loeber & Dishion, 1983). As Gottfredson (1987) states, which measure to use "depends on the question to be answered"; he continues,

> If the relative power of different devices that are developed on different populations is an issue (for which the base rates may well be different), then indices that are less sensitive to base rates would seem preferable [e.g., MCR]. However, if one wishes an estimate of the power of a particular device administered with particular decision rules on a particular population, then base rate-dependent indices will be more informative [e.g., Pearson's ϕ].

For a review of the relationships between several measures of predictive accuracy, see Tarling (1982).

Otto (1992) suggests comparing the true positive rates with the base rates: "When the true positive rate exceeds the base rate, it can be concluded that the technique is, to some extent, valid" (p. 109). S. D. Hart, Webster, and Menzies (1993) note that Otto (1992), when discussing the number of false positives, used the ratio of false positives to the sum of false positives and false negatives (what we refer to as the false positive rate; see Chapter 1); whereas Monahan (1981) used the ratio of false positives to the sum of false positives and true positives (what would be the complement of what we call the positive predictive value). The differences are obviously problematic when attempting to compare studies; S. D. Hart et al. (1993) opines that "Monahan's method (1981) gives the information of greatest interest to most reviewers, namely, the probability that a prediction of violence was incorrect" (p.

698).

Because terminology can become confusing, S. D. Hart et al. (1993) suggest that every report contain the raw data, either in the form of text or a 2×2 contingency table. The authors state this as being the minimum necessary information required, and that positive and negative predictive values should also be presented along with chance-corrected measures of accuracy such as Cohen's κ (J. Cohen, 1960) or Pearson's ϕ. Wollert (2006) also argues for the use of 2×2 tables, saying,

> Because SVP predictions classify offenders into only two groups (will recidivate or will not recidivate), scores in the alternate test range below this critical test range are considered important for identifying likely nonrecidivists. ... Once an evaluator has selected the [cutscore] he or she will use with an actuarial to select likely recidivists, he or she is able to compile a 2×2 table ... Several measures may be calculated from this simple table that are useful for evaluating the test's performance in the sample on which it was developed and for estimating its performance in another group that has a different recidivism rate from the developmental sample. (p. 58).

Mossman (1994b) disagrees, saying, "2×2 tables conflate intrinsic ability to detect future violence with the level of risk" (p. 588). Furthermore, Mossman argues that because κ, ϕ, and the positive and negative predictive values are all dependent upon the base rate, they "will not be the best indices for describing and comparing intrinsic properties of diagnostic tests" (p. 589). Mossman recommends using receiver operating characteristic (ROC) curves instead, stating that "ROC methods describe accuracy with indices of performance that are unaffected by base rates or by clinicians' biases for or against Type I or Type II prediction errors" (p. 783). An ROC curve, as described in Chapter 1 (for thorough discussions, see Green & Swets, 1966; Metz, 1978), plots the true positive rate (or sensitivity) against the false positive rate (or $1 -$ specificity). The common measure in ROC analysis is the area under the curve, or AUC; Fergusson, Fifield, and Slater (1977) show that the AUC is related to the mean cost rating. Diagnostic accuracy of predicting violent behavior refers

to an actuarial method's (or clinician's) ability to discriminate between future violent and nonviolent individuals; as Zweig and Campbell (1993) state, "[It] is the most fundamental characteristic of the test itself as a classification device" (p. 552). McFall and Treat (1999) state that ROC analysis "provides an estimate of diagnostic accuracy that is not confounded by changing cutoff values or prevalence rates" (p. 229). M. E. Rice and Harris (1995) make this point as well and demonstrate the independence of the AUC with base rates (see also Fergusson et al., 1977) in advocating the use of ROC curves (see also Swets, Dawes, & Monahan, 2000a, 2000b); the authors include correlation coefficients and odds ratios, along with the previously mentioned statistics, as measures that are dependent on base rates. For comparisons across different studies, M. E. Rice and Harris (2005) provide conversions for three different measures of effect size: AUC, Cohen's d (J. Cohen, 1969), and point-biserial correlations (see also Mossman, 2013). By the late 1990s, AUCs were, and are still the primary statistic used in establishing a diagnostic test's accuracy in predicting violence. According to their review of the literature from published articles during the years 1990–2010, Singh, Desmarais, and Van Dorn (2013) note that only three of the fifty did not report an AUC measure. Mossman (2013) suggests that AUC measures played a significant role in changing the attitude regarding the ability to predict violence with reasonable accuracy.

The use of the AUC as a measure of predictive accuracy is not without its critics. For example, Szmukler (2001) suggests that because ROCs are independent of base rates, they can be misleading. In addition, Szmukler shows that given a fixed sensitivity and specificity, the positive predictive value differs depending on the base rate. Szmukler, Everitt, and Leese (2012) state, "[T]he statistical significance of the AUC of the ROC alone offers little help when it comes to a particular patient, yet this is the statistic that is now most relied upon in the risk assessment literature" (p. 897), and that the positive predictive value is "is crucial in assessing the meaning of a positive test result for a particular patient in a particular setting" (p. 896). (These ideas are discussed thoroughly in Chapter 4.) Neller and Frederick (2013) also encourage the use of the positive predictive values and say that clinicians are misled

in RAI manuals by proportions, stating that manuals report the proportion who recidivate given a score but the real value of interest is the proportion who recidivate given the score *or higher* (i.e., the positive predictive value).

Vrieze and Grove (2008) state,

> [T]he AUC only informs one about best-case classification accuracy—accuracy that is achievable in practice only if quite restrictive assumptions are all satisfied [and i]f one or more of these assumptions is materially false, then to that extent the classification accuracy achieved may fall short of what one might expect from a published AUC figure. (p. 268)

According to the authors, these assumptions are:

> 1. the AUC is estimated without error; 2. the mathematical model relating the form of the ROC, its area, and the distributions of recidivists' and nonrecidivists' test scores is exactly valid, so that a certain equation relating the AUC to the difference between group test score distributions' means is exactly correct; 3. the recidivism base rate is estimated without error; and 4. the equation relating the forms of recidivists' and nonrecidivsts' test score distributions, their mean separation, and the recidivism base rate, is solved for the exactly optimal cutting score that minimizes classification errors. (p. 268)

They also state,

> An AUC statistic tells a researcher or clinician how well a test performs across the whole potential range of disorder (recidivism) base rates and optimum cutting scores for those base rates. It can lull the clinician into thinking that, if the AUC is suitably high, the test will perform satisfactorily in a given population, i.e., at a given base rate. This is far from necessarily so; a sufficiently high AUC only ensures that a test *can* [original emphasis] perform well in a population with a certain base rate, *if* [original emphasis] the cutting score is appropriately set. (p. 274)

In their article, Vrieze and Grove (2008) push the use and maximization of correct fractions (sum of true positives and negatives divided by the total sample) because it is a measure of a test's incremental validity over base rates (i.e., the test's predictive validity compared to simply choosing the more likely outcome). Mossman (2008) disagrees saying, "Whatever one thinks about such a policy, mental health professionals must recognize that legislatures have set a different one: identifying individuals who are 'likely to reoffend'" (p. 288).

Cook (2008) distinguishes diagnostic models from predictive (prognostic) models in that both are typically concerned with classification but the latter also incorporates a time dimension; she notes the distinction as "[p]rognostication and prediction involve estimating risk, or the probability of a future event or state. The outcome not only is unknown, but does not yet exist, distinguishing this task from diagnosis" (p. 17). Cook (2008) discusses two aspects of a model's accuracy: discrimination and calibration. In terms of violence prediction, *discrimination* is an RAI's ability to distinguish violent individuals from nonviolent ones, or different levels of risk (e.g., low, moderate, and high); *calibration* refers to the RAI's ability to estimate the risk of violence correctly. Although discrimination is a goal of both diagnostic and predictive models, calibration is a component unique to a predictive model. The discriminative component can be measured with the use of an ROC curve, but the calibration is measured by the positive and negative predictive values, values that are absent from the ROC curve. Because of this, Cook (2008) says that ROC curves are not sufficient to demonstrate the accuracy of a predictive model. Singh (2013) presents these ideas in the realm of violence prediction along with the pros and cons of many of the common statistics used (AUC and others) to assess an RAI's predictive validity. The Test Validation Study (TVS; Frederick & Bowden, 2009) provides one way of incorporating both components of predictive accuracy (see Neller & Frederick, 2013, for an application using RAIs).

Bossuyt et al. (2003) declare, "The quality of reporting of studies of diagnostic accuracy is less than optimal" (p. 7); the authors created a twenty-five item checklist they call the Standards for Reporting of Diagnostic Accuracy (STARD). Of the twenty-five items,

three are most relevant to our discussion here:

12. "Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g., 95% confidence intervals)" (p. 12)

21. "Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g., 95% confidence intervals)" (p. 15)

23. "Report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done" (p. 15)

2.14.1 Generalizability of Predictive Measures

Accuracy of the test on samples similar to the construction sample is important, but unless the measure is accurate on other populations, its use is severely limited; generalizability of RAIs should not be assumed. As Bechman (2001) puts it, "[F]ailure to scientifically cross-validate RAIs can be fatal because characteristics of offender populations can vary dramatically, and an RAI constructed on one population may not generalize, or cross over, to a different population" (p. 29). Additionally, Gottfredson (1987) warns, "There is a danger ... of overestimating the extent to which relations found in one sample can be used to explain relations in a similar sample" (p. 185).

Obviously the generalizability of predictive instruments is important and many studies have shown the instruments do generalize to other populations. For example, G. T. Harris, Rice, & Camilleri, 2004 found the VRAG to generalize to nonforensic populations (see also G. T. Harris et al., 2003). Snowden, Gray, Taylor, and MacCulloch (2007) found the VRAG and OGRS were good predictors of violent and general recidivism among UK patients, but warn that both devices over-predict recidivism. Schlager and Simourd (2007) suggest that the LSI-R could be used on Hispanic and African-American offenders. Ralston and Epperson (2013) found that the Static-99 and MnSOST-R were significant predictors of recidivism in

a juvenile sex offender sample (although they required slight revisions in scoring to account for the young ages).

Although the previous studies and others have suggested that the overall predictive accuracy of the VRAG is robust, or generalizable, across different populations, Mills, Jones, and Kroner (2005) looked specifically at the generalizability of the VRAG (and LSI-R) bins (i.e., the different probabilistic categories an individual may fall into). As the authors note, "The conclusion that the [VRAG and LSI-R] instruments are predictive of the outcome is based on aggregate analysis across the entire range of participants and scores. This approach provides limited useful information for the communication of risk within the cross-validated sample" (p. 567). In their study, the authors conclude that their results fail to "support the generalizability of the original probabilities associated with the prediction bins" for both the VRAG and LSI-R because the rates of recidivism (violent for the VRAG and violent and nonviolent for the LSI-R) within the bins were not similar to the probabilities assigned to them; the VRAG overestimated recidivism whereas the LSI-R underestimated it. They did note that the both instruments were significantly related to recidivism but suggest that "[t]he strength of the statistical relationship between an instrument's scores and outcome does not necessarily impute accuracy to the clinical application of those scores" (pp. 580).

Singh, Serper, Reinharth, and Fazel (2011) examined ten risk assessment instruments (including the COVR, HCR-20, START, VRAG, and VRS), seven of which they classified as being developed for use in mentally-ill offenders (HCR-20, START, VRAG, and VRS) and found "little direct evidence to support the use of these risk assessment tools in schizophreni[c offenders]" (p. 904).

Doren (2004a) found the rates of recidivism among most of the different Static-99 and RRASOR scores were consistent across numerous studies that were conducted in several locations around the world. However, Mossman (2006a) critiqued Doren's approach in comparing risk percentages across studies: "By directly comparing percentages of reoffenders falling in each risk category, Doren compares values that combine discriminative properties

of the risk assessment measure with the population's base rate." (p. 43). Mossman (2006a) suggests using likelihood ratios within each score to compare across studies with differing base rates; using likelihood ratios, he shows that the Static-99 and RRASOR scores were *not* consistent.

2.14.2 Authorship Bias

Hare (1998), the creator of the PCL and PCL-R, once commented,

> Researchers who obtain extremely low reliabilities for the PCL-R or its factors clearly have not used the instrument properly, do not have enough information to score the PCL-R, or have allowed the ratings to be completed by untrained or unskilled personnel (p. 107)

Murrie, Boccaccini, Johnson, and Janke (2008) found evidence of an *allegiance effect* regarding the interrater reliability of the PCL-R. The authors reviewed 43 civil commitment hearings for sexual offenders, noting that in more than half of them (23), the two legal sides significantly differed in their scores, usually in the direction of the side they represented; they called this "partisan allegiance." Boccaccini, Turner, and Murrie (2008) found similar results with the PCL-R.

In their influential article, Luborsky et al. (1999) found evidence for an allegiance effect, or *authorship bias*, in the psychotherapy literature; that is, they found effect sizes were significantly larger in studies using the author's own method or diagnostic measure. In fact, they failed to find any published result in the literature that was counter the first author's allegiance. Blair, Marcus, and Boccaccini (2008) examine whether this authorship bias exists for actuarial risk assessments. Specifically, they examine whether effect sizes (Pearson's r) reported for the Static-99, VRAG, and SORAG were significantly larger for published studies conducted by the instruments' authors than independent researchers. Their results suggest that "there is ... a pattern of allegiance in the actuarial risk assessment literature,

at least when allegiance is defined as being an author of an instrument" (p. 354). The authors conjecture several possible reasons for this including that the "instrument authors' studies incorporate certain important study design characteristics that are needed to maximize effects" (p. 354), the so-called "file drawer" effect (Rosenthal, 1979) suggesting that instrument authors are less likely to publish inconclusive results, and their own allegiance to the allegiance effect. Lilienfeld and Jones (2008) also provide three suggestions: file drawer effect, or specifically "submission bias" (authors' failure to report negative results); selective reporting where not all of the relevant results are published; and "data massaging," both defensible and indefensible treatment of data ad hoc.

In their reply to Blair et al. (2008) and Lilienfeld and Jones (2008), (G. T. Harris, Rice, & Quinsey, 2010, authors of the VRAG and SORAG instruments) performed their own analysis but using AUC measures rather than Pearson's r. They point out several flaws in Blair et al.'s (2008) work; specifically, G. T. Harris et al. disagree with Blair et al.'s decision to only include published studies. They included all unpublished studies using the VRAG and SORAG (none by the authors themselves); they conclude "that examining all currently available data yields no evidence of an allegiance effect, and therefore, no basis in our work for a commentary (Lilienfeld & Jones, 2008) on the topic" (p. 86).

Singh, Grann, and Fazel (2013) examine authorship bias in nine different actuarial tools from published journals, conference presentations, government reports, and theses and, using odds ratios (the ratio between the odds of a true positive and the odds of a false positive), found "[e]vidence of a significant authorship effect ... specifically to risk assessment studies published in peer-reviewed journals" (p. 6). Possibly more disconcerting was the authors' finding of consistent failure to report conflict of interest statements:

> Six of the 16 journals in which the studies appeared requested in their Instructions for Authors that any financial or non-financial conflicts of interest be disclosed. None of the 25 studies where a tool designer or translator was the author of an investigation of that instrument's predictive validity contained such a disclosure.

2.15 Admissibility of Actuarial Methods

Based on the numerous failed attempts to challenge the constitutionality of preventive detention despite disagreements among many academics, it appears that the courts have expressed their opinion. But what about the admissibility of the methods making those predictions? Despite a plethora of studies indicating the inability of clinicians to make predictions regarding future dangerous behavior, the ruling in *Barefoot v. Estelle* (1983) made it clear that because it was not impossible to predict future behavior, it was admissible. As Janus and Meehl (1997) state, "As a legal matter, prediction is not, in all of its forms and for all purposes, so inaccurate as to violate the due process clause" (p. 36). Following the ruling in *Tarasoff v. Regents of the University of California* (1976), as Monahan (1996) put it, "Liability, rather than constitutionality, [wa]s the concern that motivate[d] interest in the prediction of violence in the mid-1990s" (p. 111).

L. Walker and Monahan (1988) provide a thoughtful review of the use of social science research in the legal system. Berlin et al. (2003) suggest that assessment instruments, such as the RRASOR, MnSOST-R, and Static-99 should be used to screen potentially violent individuals but not used as evidence for an individual's risk in civil commitment hearings. Mossman et al. (2011) provides a detailed review of state laws addressing risk of dangerousness and the role of RAIs in determining this risk. Vitacco, Erickson, Kurus, and Apple (2012) review 46 legal cases that involved the use VRAG or HCR-20, both instruments designed to predict or assess risk of violence; the authors found that most cases where the instruments are used are SVP cases involving commitment or release, which the authors note is justifiable given the research. The authors do caution that risk assessment instruments should be used "in a manner that is empirically justified and consistent with research," citing cases involving the death penalty and juveniles being tried as adults as situations where risk assessment measures may not be appropriate.

Doren (2000) discusses the admissibility of actuarial instruments used to make decisions regarding civil commitment of sexual offenders. He addresses several issues he claims were commonly made, the most relevant being that the instruments possess only moderate accuracy. Doren states that civil commitment laws require a degree of likelihood that an individual sexual offender will recidivate and not a determination of recidivism (i.e., a prediction). He also suggests that statistics such as the AUC may not be appropriate in determining whether a sexual offender's recidivism risk is beyond the legal threshold; rather, he suggests the use of confidence intervals to provide more meaningful distinctions among the recidivism risk levels. In their reply to Doren (2000), Otto and Petrila (2002) state that it is often assumed by the lawyers and judges of the court that predictions regarding recidivism are in fact being made about a sexual offender, providing excerpts of court transcripts offering such suggestive language. Doren (2000) also addresses the issue that actuarial instruments may not meet psychological test standards, by stating that the instruments are "clearly not psychological tests" (p. 66). He also makes the peculiar comparison of insurance actuarial methods with those used for sexual recidivism (similarly done in Skeem & Monahan, 2011). He states, "successful use in other fields [e.g., insurance] demonstrates that the failure to meet standards for psychological testing does not summarily negate the value or usefulness of data derived from actuarial instruments in assessing risk factors" (p. 78). Otto and Petrila (2002) point out that actuarial instruments for sexual recidivism have (a) unknown inter-rater reliabilities, (b) unknown rates of measurement error, (c) lack cross-validation, and (d) lack minimally acceptable test manuals. Otto and Petrila note that some courts have characterized the instruments as "tests" and state, "If these are not 'tests,' those providing testimony based on them should explain to the court how they differ from instruments meeting test standards and how those differences diminish the credibility of the testimony based on them" (p. 14).

Janus and Prentky (2003) state,

> Although preventive detention would be legally and ethically problematic even with

> perfect knowledge about the future, the imperfection of risk assessment exacerbates constitutional and ethical concerns because it raises the likelihood that non-recidivists and low-risk individuals will be among the group suffering long-term loss of liberty. The same is true for the more utilitarian concerns about resource allocation and efficacy. The central justification for spending huge sums of money on SVP programs is that the "most dangerous" offenders are incapacitated. Public policy is not well served if, because of inaccurate assessment of risk, extraordinary resources are devoted to the ordinarily dangerous (Janus, 2003). (p. 1448)

Janus and Meehl (1997) examine the evidence used in court to estimate the standard of proof for predicting future dangerousness used in the legal setting. As the authors argue, civilly committing sexual offenders based on dangerousness implies,

> (a) the probability of dangerousness is susceptible of measure, (b) there is a way to discriminate between predictions of higher and lower probability, (c) there are standards that allow commitments based on the former while excluding confinement based on the latter, and (d) these standards are, in fact, enforced. (p. 38)

Janus and Meehl's paper focuses on these last two points; they point out that errors in decisions (commitment or release) occur because the decision is dichotomous and based on continuous data that is "imperfectly perceived" (p. 38). The causes of these erroneous decisions, they argue, come from two sources: imperfect perception from the decision maker and the probabilistic nature of risk (e.g., the event where a truly low-risk individual engages in a violent act). Janus and Meehl (1997) note that the courts use vague terms such as *highly likely* to describe an offender's risk of recidivism without having a standard for quantifying this risk. By using evidence from previous court cases, the authors attempt to determine a quantification of said vagaries. As an example of how judicial systems may (mis-)quantify such terms, they cite the case of *In re Young* (Wash. 1993) involving a Washington State sex offender, Andre Young, serving an indefinite involuntary commitment term:

> [T]he Washington Supreme Court quoted a prominent sex offender scholar: "'using theoretically relevant and empirically tested predictors, predictive accuracy [of sexual recidivism] can realistically be expected to be in the 80% range'." The Court used this quote to support its assertion that the "likelihood of reoffense is extremely high" among those subject to commitment. (p. 40).

The authors note this incorrect equivalence of witness accuracy and probability of a correct identification as a common heuristic bias. This bias is, in fact, an example of what is known as the *Prosecutor's Fallacy*.

The Prosecutor's Fallacy (also called the *Inversion Fallacy*) was first coined by W. C. Thompson and Schumann (1987). This fallacy involves misinterpreting conditional probabilities; for example, misinterpreting the sensitivity as the positive predictive value. The statement "predictive accuracy" quoted in *In re Young* (Wash. 1993) was given by Dr. Vernon Quinsey, one of the authors who developed the VRAG, and can be found in *Review of Sexual Predator Program: Community Protection Research Project* (Washington State Institute for Public Policy, 1992). Dr. Quinsey was presumably referring to the sensitivity of the test, meaning the number of recidivists correctly predicted to recidivate (this is the probability that an individual attains a given score on the actuarial measure given that the individual will truly recidivate). The mistakenly-assumed equivalent statement, made by the Court, was that "likelihood of reoffense is extremely high." This quantification is measured by a different conditional probability, namely the positive predictive value (i.e., the probability an individual will recidivate given the individual attains a given score on the actuarial method). This probability, as we will see in several subsequent chapters, can be vastly different than the sensitivity of the test, particularly when the base rate for violence is low. Dr. Quinsey could have instead been referring to the overall accuracy of the actuarial method, as implied by Janus and Meehl (1997, see p. 44, footnote 59). The accuracy, as we will see in the next section, depends on conditional and prior probabilities. Similarly, this can vary dramatically from the positive predictive value depending on the prior probabilities. In ad-

dition to misinterpreting conditional probabilities, the incorrect assumption of independence can play a scathing role in the legal system (e.g., see Tribe, 1971, for a thorough discussion on the use of mathematics in the judicial system; also, see Hubert & Wainer, 2012, and Schneps & Colmez, 2013, for numerous examples of the misuse of probability, statistics, and mathematics in the legal system).

Janus and Meehl (1997) used probabilities of recidivism of 50% and 75% to quantify the terms "likely" and "highly likely," respectively. The authors defined their "commitment standard" as the ratio of correct commitments to total commitments, claiming it to be "the probability that a person actually committed by the courts would have been a recidivist had it not been for the commitment" (p. 43). The authors used two models to emulate the process of civil commitment in Minnesota sex offender commitment cases. The authors' research, based on empirical base rates and overall accuracy for predicting recidivism, led them to,

(a) conclude, using Model 1, that "a standard of commitment of 50% appears achievable but only under favorable conditions" and "rule out the possibility that sex offender commitment courts are using a standard of commitment that approximates 75%" (p. 55);

(b) and, using Model 2, that "Minnesota courts never attain a 50% probability of recidivism standard of commitment" (p. 58).

Mossman (2006b) says,

> [W]e may need to abandon the hope that more accurate methods of predicting violence or assessing patients' level of violence risk will prove useful to practicing clinicians. Abandoning this hope may be cause for initial disappointment. But if courts and my mental health colleagues agree with my conclusion, abandoning the hope for useful risk assessments will ultimately liberate us from obligations that we

> cannot carry out rationally, and will allow us to refocus our attention on treating patients (p. 530)

2.15.1 The Federal Rules of Evidence

In *Frye v. United States* (D.C. Cir. 1923), the use of a polygraph test as acceptable scientific evidence was challenged. The Court stated,

> We think the systolic blood pressure deception test has not yet gained such standing and scientific recognition among physiological and psychological authorities as would justify the courts in admitting expert testimony deduced from the discovery, development, and experiments thus far made. (p. 1014)

In addition, the Court noted the difficulty in determining what constituted scientifically-sound evidence:

> Just when a scientific principle or discovery crosses the line between the experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone the evidential force of the principle must be recognized, and while courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs.

The *Frye* standard refers to the admissibility of scientific evidence; regardless of the source of the evidence (e.g., from a so-called expert), the evidence is only admissible if the method generating the evidence (e.g., an actuarial method) is considered reliable among the scientific community.

In 1975, the United States federal courts established a code of evidence law, since revised, entitled the Federal Rules of Evidence (FRE). The purpose of the FRE is "to secure fairness in administration, elimination of unjustifiable expense and delay, and promotion of

growth and development of the law of evidence to the end that the truth may be ascertained and proceedings justly determined" (Federal Rules of Evidence, 2013, p. 1). Rule 702, regarding testimony made by expert witnesses, states:

> A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:
>
> (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
>
> (b) the testimony is based on sufficient facts or data;
>
> (c) the testimony is the product of reliable principles and methods; and
>
> (d) the expert has reliably applied the principles and methods to the facts of the case. (p. 30)

United States Federal Judge Jack Weinstein comments on Rule 702, and others, as well as the use of probability and statistics in the legal setting in his article *Litigation and Statistics* (1988). Regarding Rule 702(b) he states, "The breadth of the 'assistance to the trier of fact' standard has encouraged courts to adopt a balancing test in determining the admissibility of expert testimony" (p. 287). Regarding expert testimony in general, he notes,

> Unfortunately, we have concluded that an expert can be found to testify to the truth of almost any factual theory, no matter how frivolous, thus validating the case sufficiently to avoid summary judgment and forcing the matter to trial. At the trial itself, an expert's testimony can be used to obfuscate what would otherwise be a simple case. The most tenuous factual bases are sufficient to produce firm opinions in some experts to a high degree of "probability." Juries and judges can be, and sometimes are, misled by such experts for hire. (p. 289)

Slobogin (1984) argues that if clinicians testify they need to limit their statements to within the realm of their capabilities, excluding predictions of individual dangerousness:

> Unfortunately, where clinical predictions are concerned, even those individuals who have acquired some specialized knowledge about dangerousness often provide testimony exceeding the boundaries of their knowledge. In particular, clinical predictions that an individual is dangerous or "likely to be violent," although presently the mainstay of commitment and sentencing proceedings, go beyond the limited scope of current expertise. There simply are no clinical theories that permit such bald assertions. (p. 130)

He then states, "The Federal Rules of Evidence do permit an expert to address the 'ultimate issue' to be decided by the trier of fact [e.g., the jury]" (p. 130). The author suggests that clinical predictions of dangerousness should only be permitted if the defendant chooses to use such testimony or if the predictions are actuarial, unless in a civil commitment setting "because the nature of the dangerousness inquiry that it requires is fundamentally different from the prediction process [encountered in the sentencing and criminal commitment contexts]" (p. 171).

In *People v. Murtishaw* (1981), the defendant, David Murtishaw, having been convicted on three counts of first degree murder and one count of assault with intent to commit murder, went to the penalty phase of his trial in California. Dr. Ronald Siegel, a psychopharmacologist, testified,

> [The] defendant in a prison setting "will continue to be a violent assaultive and combative individual ... [a]nd that he may become not only assaultive and violent, but he could show the same types of homicidal tendencies that he has shown in the past, with no ability to morally or physically constrain himself to the demands of the environment in which he finds himself." (p. 767)

The California Supreme Court, in its opinion, states that Dr. Siegel's testimony should not have been permitted because:

> (1) expert predictions that persons will commit future acts of violence are unreliable, and frequently erroneous; (2) forecasts of future violence have little relevance to

> any of the factors which the jury must consider in determining whether to impose the death penalty; [and](3) such forecasts, despite their unreliability and doubtful relevance, may be extremely prejudicial to the defendant. (p. 767)

The Court particularly disapproved of predictions of dangerousness in capital punishment cases: "In short, evidence which is barely reliable enough to justify a civil judgment or a limited commitment is not reliable enough to utilize in determining whether a man should be executed" (p. 771).

The United States Supreme Court determined a standard for admissible scientific data and expert testimony used in the United States federal court system in *Daubert v. Merrel Dow Pharmaceuticals, Inc.* (1993). Two major consequences of the ruling were the suspension of the *Frye* Standard by the FRE, and the establishment of the trial judge as a "gatekeeper" for allowing evidence that is determined relevant and scientifically valid. Several relevant points regarding the admissibility of a theory or technique were laid forth, requiring that they be:

(a) falsifiable, refutable, or testable;

(b) subjected to peer review and publication;

(c) have a known or potential error rate; and

(d) reliable and relevant among the appropriate scientific community.

Two other landmark cases, *General Electric Co. v. Joiner* (1997) and *Kumho Tire Co. v. Carmichael* (1997), together with *Daubert v. Merrel Dow Pharmaceuticals, Inc.*, are known as the *Daubert* trilogy; they set forth what has been called the *Daubert* standard (i.e., the four requirements stated above). Rule 702 from the FRE, as revised to its current 2013 state and provided earlier, follows the rulings in the *Daubert* trilogy. Most state judicial systems have elected to follow the *Daubert* standard, although some states still follow the *Frye* standard. Scherr (2003) notes that "since Daubert, every appellate court to have

reviewed the question has admitted expert predictions" (p. 59) and that "[j]udicial opinion, split on virtually every other form of behavioral or psychic expertise, has so far unanimously accepted predictive expertise in civil commitments" (p. 61). (For a thorough review of the FRE and *Daubert* trilogy and their relationship to predictions of dangerous behavior, see Scherr, 2003; Slobogin, 2006.)

In *State v. Randall* (1995), the Supreme Court of Wisconsin summarized the findings of the United States Supreme Court, many that have been discussed in this chapter:

> [T]he Supreme Court [of the United States] has found the following to be constitutionally permissible: civil and criminal insanity acquittees may be treated differently with regard to the burden of proof required for the initial commitment—commitment following an insanity acquittal may be based on a preponderance of the evidence, whereas in a civil commitment the state must establish its burden of proof by clear and convincing evidence; automatic commitment following an insanity acquittal does not violate due process; the length of commitment for both civil and criminal committees may be indefinite; the term of commitment for an insanity acquittee may exceed the length of the maximum sentence the acquittee could have been subjected to had a sentence been imposed; and, following the initial commitment of an insanity acquittee, the burden of proof at a subsequent hearing for reexamination and release may be borne by the acquittee. (p. 821)

Noticeably absent from the above summary is the use of testimony (actuarial or clinical) regarding dangerousness in capital punishment phases; in *Flores v. Johnson* (2000), United States Court of Appeals Fifth Circuit Judge Emilio Garza offered his opinion on the matter:

> On the basis of any evidence thus far presented to a court, it appears that the use of psychiatric evidence to predict a murderer's "future dangerousness" fails all five *Daubert* factors. First, "testing" of these theories has never truly been done ... Second, as is clear from a review of the literature in the field, peer review of

> individual predictions is rare, and peer review of making such predictions in general has been uniformly negative. ... Third, the rate of error, at a minimum, is fifty percent, meaning such predictions are wrong at least half of the time. ... Fourth, standards controlling the operation of the technique are nonexistent. ... Overall, the theory that scientific reliability underlies predictions of future dangerousness has been uniformly rejected by the scientific community absent those individuals who routinely testify to, and profit from, predictions of dangerousness. (pp. 464–465, citations omitted)

2.15.2 *State of New Hampshire v. William Ploof* (2009)

In 1998, William Ploof was sentenced to four to ten years for felonious sexual assault at the New Hampshire State Prison, but after his maximum release date in 2007, he continued to be held in prison (*State of New Hampshire v. William Ploof*, 2009). According to the state's SVP laws, a multidisciplinary team (MDT) established by the New Hampshire Department of Health and Human Services (DHHS) is required to determine whether persons convicted of sexually violent offenses and eligible for prison release, such as Mr. Ploof, meet the definition of a "sexually violent predator" (N.H. Stats., 2014b, Ch. 135-E:3.I). If so, the county attorney or attorney general can file a petition with the state's superior court to argue the case against the plaintiff. In New Hampshire, a sexually violent predator is defined as a person who "has been convicted of a sexually violent offense" and "suffers from a mental abnormality or personality disorder that makes the person likely to engage in acts of sexual violence if not confined in a secure facility for long-term control, care, and treatment" (N.H. Stats., 2014b, Ch. 135-E:2.XII (a)–(b)). In the case of Mr. Ploof, the MDT consisted of two psychologists and a DHHS employee; the Static-99 was used by the MDT to determine Mr. Ploof's risk of recidivism. In *State of New Hampshire v. William Ploof* (2009), the admissibility of the Static-99 was essentially put on trial, in particular the Static-99R; although it is never referred to as the Static-99R in the court transcripts (only the Static-99), the revised

norms of the Static-99 are continually referenced.

In July 2004, the New Hampshire state courts set forth reliability standards of experts, derived from the *Daubert* standard. In Section 516:29-a of the New Hampshire Revised Statutes Annotated (RSA; N.H. Stats., 2014a), requirements for the testimony of expert witnesses are as follows:

I. A witness shall not be allowed to offer expert testimony unless the court finds:

   (a) Such testimony is based upon sufficient facts or data;

   (b) Such testimony is the product of reliable principles and methods; and

   (c) The witness has applied the principles and methods reliably to the facts of the case.

II. (a) In evaluating the basis for proffered expert testimony, the court shall consider, if appropriate to the circumstances, whether the expert's opinions were supported by theories or techniques that:

   (1) Have been or can be tested;

   (2) Have been subjected to peer review and publication;

   (3) Have a known or potential rate of error; and

   (4) Are generally accepted in the appropriate scientific literature.

   (b) In making its findings, the court may consider other factors specific to the proffered testimony.

According to the court transcripts for *State of New Hampshire v. William Ploof* (2009), to comply with the RSA requirements stated above, the expert testimony "does not have to reliably *predict* [original emphasis] whether a particular individual will or will not recidivate ... Rather ... [it has to] reliably differentiate[] between those 'likely' to recidivate and those with a lesser recidivism risk" (*State of New Hampshire v. William Ploof*, 2009, pp. 6–7). In the court hearings, it was testified by Dr. Amy Phenix, a clinical psychologist and "contributor to the Static-99 Coding Rules Revised" (Static-99, 2013),

and concluded by the Court that the "Static-99 can be described as having 'moderate' predictive accuracy" (*State of New Hampshire v. William Ploof*, 2009, p. 17); this was based on validation studies of the Static-99 finding AUC values as high as .71. The AUC of the Static-99, according to testimony given by Dr. Phenix, is .68; for the new norms .665 (at the time of the hearing, the new norms had not been cross-validated). She also testified and led the Court to conclude that the "Static-99 has an accuracy rate that is significantly better than chance" (p. 15), "the use of multiple actuarial tools d[oes] not increase the predictive accuracy of the assessment" (pp. 13–14), and "the use of empirically guided clinical judgment to adjust the results may decrease the overall predictive accuracy of a particular assessment" (p. 30). However, Dr. Phenix also led the Court to conclude that "an evaluator should not base his or her conclusions about an individual's recidivism risk solely on the Static-99[R]" (*State of New Hampshire v. William Ploof*, 2009, p. 35).

It is recommended by Helmus et al. (2009) that once a range is determined by the Static-99R, a judgment be made where the individual falls within this range. The Court ruled this clinical judgment fails to meet the requirement of the RSA 516:29-a I(b), but did eventually conclude that the Static-99R is admissible in the court of law. This decision appeared to be dictated by the accuracy of the Static-99R, quantified by the AUC values.

2.15.3 Probabilities and the Law

In the memorandum from *United States v. Fatico* (1978), Judge Jack Weinstein discussed the continuum of burden of proof; the terms associated with the continuum were mentioned several times throughout the chapter, but will be discussed here in further detail.

The constitutionally-mandated standard for criminal conviction in the United States is *guilty beyond a reasonable doubt.* As Tribe (1971) puts it,

> The jury is charged that any "reasonable doubt," of whatever magnitude, must be resolved in favor of the accused. ... for the jury to announce that it is prepared to convict the defendant in the face of an acknowledged and numerically measurable

> doubt as to his guilt is to tell the accused that those who judge him find it preferable to accept the resulting risk of his unjust conviction than to reduce that risk by demanding any further or more convincing proof of his guilt. (p. 1374).

Given the evidence (E), an individual is guilty (G) when $P_j(G|E) \geq b_j$ where b_j is the probability threshold assigned by juror j. Fienberg (1989) notes that b_j should be quite close to 1 when the burden of proof is beyond a reasonable doubt. He also notes, as does Tribe, that $b_j \neq 1$, for if $b_j = 1$ this would correspond to "proof beyond any doubt" (p. 199). Fienberg (1989, see also Fienberg & Kadane, 1983) also note that "innocent until proven guilty" might imply that $P_j(G) = 0$; but from a Bayesian perspective this would mean that $P_j(G|E) = 0$ for all E.

In civil cases, the standards are generally less strict than in criminal cases; that is, the juror or judge follows the notion of *preponderance of evidence*—more probable than not—where the law "seeks to minimize the probability of error" (*United States v. Fatico*, 1978, p. 403). Fienberg and Kadane (1983) state this corresponds to a Bayesian requirement that $P_j(T|E) \geq d_j$ where T (the authors use P) is the "event that the plaintiff's version of the issue in dispute is correct" (p. 200) and $b_j > d_j \geq 1/2$. Fienberg and Kadane (1983) note that legal theory requires $P_j(T) \geq 1/2$ but it is often assumed that the theory suggests $P_j(T) = 1/2$; they state, "it would be impossible to find a jury for whom $P_j([T]) = 0.5$, exactly for each juror j" (p. 93).

According to Weinstein, civil proceedings use the standard of *clear and convincing evidence* when "moral turpitude is implied" (p. 404). As noted earlier, this is often the standard for civil commitment (*Addington v. Texas*, 1979, i.e.,); it requires a burden of proof $P_j(C|E) \geq c_j$ where $b_j > c_j > d_j$. Weinstein also included *clear, unequivocal, and convincing evidence* as a more demanding burden of proof than clear and convincing, but less than beyond a reasonable doubt.

Simon and Mahan's (1971) research found that both university students and potential jurors serving on jury duty service were less likely to find a defendant guilty if first asked

to assign a probability of guilt. The authors also asked judges, jurors, and university students to provide a probability quantifying proof beyond a reasonable doubt. (Note that the probabilities were based on a scale ranging from 0–10 by increments of 0.5, but are adjusted here to a 0–1 range; see Figure 2 in Simon & Mahan, 1971 for more details.) The authors found that the three groups had similar ideas of what beyond a reasonable doubt meant probabilistically; the median for all four groups was in the range $(.85, .90]$; eight respondents (three students and five jurors) gave a probability less than .50. In contrast, when asked to provide a probability estimate for proof by preponderance of evidence, judges' median estimate was in the range $(.50, .55]$, as Fienberg and Kadane (1983) suggests; however, for jurors and students the median was $(.75, .80]$ and $(.70, .75]$, respectively. As the authors state,

> [F]or these lay groups, the difference between the criminal (beyond a reasonable doubt) and civil (by a preponderance of the evidence) standards are much less than they are for the judges. The judges make a much sharper distinction between the criminal and civil standards. (p. 325)

In discussing Simon and Mahan's research, Underwood (1977) believes the results suggest that "for most people the distinction [i]s clear" (p. 1311). She continues,

> There is some evidence, then, that factfinders can distinguish among degrees of belief, and that rules about the burden of persuasion affect the outcome of cases. It is at least plausible, therefore, that the requirement of proof beyond a reasonable doubt serves the purposes attributed to it. (p. 1311)

Underwood's assertion implies that, although the probabilistic definitions may not be correct, the fact that the two burdens are distinguished is important.

Weinstein surveyed ten judges in the Eastern District of New York asking for probabilistic assessments of the four burdens of proof (*United States v. Fatico*, 1978); the results are reproduced in Table 2.1. Given the varying probabilistic assignments, an interesting contradiction arises if one considers the results from Monahan and Silver (2003) discussed earlier.

Recall that their results found that judges were willing to civilly commit an individual given an estimated probability of violence of .26 or greater. In other words, given the evidence provided by the COVR the average judge in their study would require $P_j(C|E) \geq .26$, where C is the decision to civilly commit an individual and E is the score of the COVR. These results seem to contradict the belief that $b_j > c_j > d_j > 1/2$, where $c_j = .26$ is the probability measure assigned to clear and convincing evidence, the typical burden of proof in a civil commitment case. Based on the results in Weinstein's table, the threshold for the COVR score should be either the high-risk or very high-risk category (corresponding to estimated probabilities of .56 or .76, respectively). The only true difference is that the COVR score (the evidence) is for an event that *may* occur, whereas the earlier discussion in this section considers evidence for an event that did occur.

Judge	*Preponderance*	*Clear and Convincing*	*Clear, Unequivocal and Convincing*	*Beyond a Reasonable Doubt*
1	50+%	60–70%	65–75%	80%
2	50+%	67%	70%	76%
3	50+%	60%	70%	85%
4	51%	65%	67%	90%
5	50+%	—	—	90%
6	50+%	70+%	70+%	85%
7	50+%	70+%	80+%	95%
8	50.1	75%	75%	85%
9	50+%	60%	90%	85%
10	51%	—	—	—

Table 2.1: A survey of ten judges from the Eastern District of New York assessing probabilistic values to the four burdens of proof (see table on page 410 in *United States v. Fatico*, 1978).

2.16 Prediction Hits the Streets

The following plot summary describes the 2002 movie *Minority Report*:

> In the year 2054 A.D. crime is virtually eliminated from Washington D.C. thanks to an elite law enforcing squad "Precrime." They use three gifted humans (called "Pre-Cogs") with special powers to see into the future and predict crimes beforehand. (Soumitra, 2014)

In the movie, John Anderton (played by Tom Cruise) arrests Howard Marks (played by Arye Gross), saying,

> Mr. Marks, by mandate of the District of Columbia Precrime Division, I'm placing you under arrest for the future murder of Sarah Marks and Donald Dubin that was to take place today, April 22 at 0800 hours and four minutes. (de Bont, Curtis, Molen, Parkes, & Spielberg, 2002)

Predictive policing is the real world version of Minority Report's Pre-Cogs. PredPol is a predictive policing software program developed by scholars at several California universities; according to its website (http://www.predpol.com), PredPol is "[b]uilt upon the same underlying technology used to predict aftershocks following earthquakes" and "[p]redicted twice as much crime as experienced crime analysts in 6 month randomized control trials" (PredPol, 2014). PredPol is just one of many new softwares being developed to predict crime and, according to Sengupta (2013), this is "a large and fast-growing market" (para. 12). Much of the discussion in the media has touted predictive policing as being able to prevent crime at lower costs (e.g., see Goode, 2011; J. Rubin, 2010), and lower crime rates in cities using predictive policing are often cited as proof of its success. But Sengupta also points to some of the issues behind predictive policing, particular a self-fulfilling bias: "[A]n area with historically high rates of crime gets greater police attention, which results in more arrests, which in turn the algorithm uses to deem that neighborhood an area where crime is more likely to occur" (para. 9). And Vlahos (2011) says, "The dirty secret of the futuristic approach [i.e., predictive policing] ... is that nobody knows for certain that it works" (p. 64).

In an article entitled *Misfortune Teller* appearing in *The Atlantic*, Nadya Labi discusses the realization of predicting crime before it occurs, focusing on Professor Richard Berk: "[He] likes to think he knows what criminals will do—even before they know," the author says (Labi, 2011, para. 1). Labi describes how Berk's methods using decision trees have been implemented in the state of Pennsylvania, methods that he says, when compared to older ones, are "like comparing a Ford Focus to a Ferrari" (para. 6).

And with the use of predictive policing, the clinical-versus-statistical-prediction debate may again rear its ugly head:

> Computers are better at flagging statistical trends, but cops still have to interpret them, says Lt. Col. Howell Starnes of the [Memphis Police Department]. "Until you get that street officer who knows his ward, you won't know what's *causing* [original emphasis] the crime," he says. "That's what you've got to look at. Not that you've got a problem—what's causing the problem." (Vlahos, 2011, p. 65)

2.16.1 Stop-and-Frisk

In 1968, the United States Supreme Court ruled that confrontation involving a police officer investigating a citizen under suspicion, the police officer has the right to "stop-and-frisk" (*Terry v. Ohio*, 1968). The defense in the case argued unsuccessfully that this was unconstitutional, citing that it violated the Fourth Amendment, which states,

> The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized. (U.S. Const. amend. IV, n.d.)

In 1964, New York State amended its Code of Criminal Procedure law (New York Criminal Procedure, 2014) to allow the police to stop and question any person suspicious of criminal activity (Ronayne, 1964). Harcourt (2008) notes,

> [T]he implementation of a targeted policing strategy focused on increased stop-and-frisk searches on the streets of New York City in the 1990s resulted in disproportionate searches of African-American and Latino citizens, as well as a sharp rise in the number of civilian complaints of police misconduct, including brutality (pp. 30–31)

In 2010, New York's law gained national attention after former New York City Policeman Adrian Schoolcraft released audio recordings revealing, among other things, specified stop-and-frisk quotas set forth by the NYPD (Rayman, 2010). Shortly after, several civil rights groups filed a class-action lawsuit targeting the stop-and-frisk program. The New York City Liberties Union reported that "[n]early nine out of 10 stopped-and-frisked New Yorkers have been completely innocent"; that every year since 2003 over half of those stopped were black, around one-third were Latino, and about ten percent were white; and that the number of stop-and-frisks increased every year from 2002 (97,296 reported cases) to 2011, except in 2007 (New York Civil Liberties Union, 2014). Facing criticism for his city's police department, then-mayor Michael Bloomberg stated, "We are not going to walk away from a strategy that we know saves lives" (K. Taylor, 2012).

In *Floyd v. The City of New York* (2013a), the stop-and-frisk law, as carried out by the NYPD, was ruled unconstitutional by Judge Shira Sheindlin; she stated, "Without a system to ensure that stops are justified, such pressure [to make a certain number of stops] is a predictable formula for producing unconstitutional stops" (p. 10). She also noted,

> The City and its highest officials believe that blacks and Hispanics should be stopped at the same rate as their proportion of the local criminal suspect population. But this reasoning is flawed because the stopped population is overwhelmingly innocent—not criminal. There is no basis for assuming that an innocent population shares the same characteristics as the criminal suspect population in the same area. (pp. 8–9)

She did not rule the stop-and-frisk program to be terminated, but rather that it be "carried out in a manner that protects the rights and liberties of all New Yorkers" (*Floyd v. The City of New York*, 2013b).

2.17 Neuroprediction of Violence

Neurocriminology is a neurobiological field of criminology (i.e., the study of brain- and cognitive-related risk factors in crime). In his book, *Anatomy of Violence*, Adrian Raine tells the story of a middle-aged man who develops a tumor and begins displaying pedophilia-like symptoms that include fondling his step-daughter and collecting pornography of minors. After failing treatment, the man is to be sent to prison; just prior, the tumor is discovered and removed. After the removal he appears to return to his normal self. Raine suggests that the case is "as close as one can get to demonstrating a *causal* [original emphasis] link between brain dysfunction and deviant behavior" (p. 305). The author examines biological, physiological, and neurological factors that may be predictive of violence, including vitamin deficiency at an early age, low-resting heart rate, and an abnormal brain structure (Raine, 2013). Interestingly, many of these predictive factors that Raine cites can be found in himself (P. Bloom, 2013).

Neuroprediction uses neuroscientific results (e.g., different brain compositions) to make predictions. This area of research has garnered recent attention in predicting future violence (Nadelhoffer et al., 2012; Nadelhoffer & Sinnott-Armstrong, 2012); however, it is not novel (e.g., Italian criminologist Cesare Lombroso believed criminals could be distinguished from noncriminals based on biological features, Lombroso, 1911). Aharoni et al. (2013) found that hemodynamic activity in the anterior cingulate cortex was a significant predictor of recidivism in a modest-sized sample ($N = 96$) of adult male offenders.

Because this is a relatively new area for violence prediction and not a lot of research exists, it is likely to be one that receives considerable attention in the years to come (Lam-

parello, 2010; Looney, 2009; Monahan, 2013).

2.18 Where to Next?

Melamed, Bauer, Kalian, Rosca, and Mester (2011) suggests that violence risk assessment can and should be used before issuing firearms licenses (see also Consortium for risk-based firearm policy, 2013); one may imagine how this requirement would hold up against the Second Amendment in the United States. Several laws across the United States have banned—often over a specified period of time—gun possession to those who have been involuntarily committed or have a mental disorder and a history of violence (Luo & McIntire, 2013). When guns are confiscated from persons with mental disorders they are often returned; some jurisdictions require that mentally-ill persons be cleared by a mental health professional, and this may be in the form of a "doctor's note certifying that the gun owner is no longer a danger" (Luo & McIntire, 2013, para. 55). In Maine, possession of firearms are prohibited if a person has been "[c]ommitted involuntarily to a hospital" (Maine Criminal Procedure, 2014, 1.E(1)); however, under certain circumstances (e.g., emergency involuntary commitment) the law has been deemed unconstitutional (*United States v. Rehlander*, 2012). The Maryland Court of Special Appeals arrived at a similar ruling in *Furda v. State* (2010):

> [When] a judge approves an ex parte petition for an emergency, involuntary mental health evaluation is hardly the equivalent of a commitment. In our view, that would be akin to suggesting that an arrest warrant approved by a judge is the equivalent of a conviction. (p. 888)

Monahan (2012) discusses risk assessment of terrorism, presenting risk factors, approaches and considerations, and methodological challenges, particularly noting the difficulty of validating such a predictive measure. The United States' Department of Homeland Security (DHS) developed the Future Attribute Screening Technology (FAST) to "rapidly identify suspicious behavior indicators to provide real-time decision support to security and

law enforcement personnel" (U.S. Department of Homeland Security, 2008, p. 2). FAST was designed to measure heart rate and respiration, track position and gaze of eyes and measure pupil diameters, detect thermal changes in the facial skin, and provide detailed images of the face and body and audio analysis determining pitch changes. Because a terrorist act is an extremely rare event, the number of false positives is an immediate concern (Furnas, 2012; Weinberger, 2012). The DHS has attempted to measure the accuracy of FAST by "instructing some people passing through the system to carry out a 'disruptive act'" (para. 6), reporting an accuracy of around 70% (Weinberger, 2012). As Furnas (2012) notes, this means that FAST "would produce false positives at an abysmal rate" (para. 6).

2.19 Conclusion

Nearly a year after the Newtown shooting, and several mass murders since, the state of Connecticut released its report on the Newtown shootings (State of Connecticut Division of Criminal Justice, 2013). The report states,

> It is known that the shooter had significant mental health issues that affected his ability to live a normal life and to interact with others, even those to whom he should have been close. As an adult he did not recognize or help himself deal with those issues. What contribution this made to the shootings, if any, is unknown as those mental health professionals who saw him did not see anything that would have predicted his future behavior. (p. 3)

This was reiterated later in the report: "It is important to note that it is unknown, what contribution, if any, the shooter's mental health issues made to his attack on [Sandy Hook Elementary School]" (p. 35). The report concludes,

> [Adam Lanza's] mental status is no defense to his conduct as the evidence shows he knew his conduct to be against the law. He had the ability to control his behavior to obtain the results he wanted, including his own death. ... The existence of an

> extreme emotional disturbance for which there is a reasonable explanation or excuse is also not present in this case. (p. 42)

The report makes it clear that although Lanza may have been mentally ill, there was no reason to suspect it was a cause of his violent behavior.

The history of violence prediction is a long and controversial one and will continue to be. As computational power increases and "big data" continues to be popular, it may appear that prediction becomes easier. But, as we have seen, many of the issues that initially existed continue to exist today. As Mayer-Schönberger and Cukier (2013) state,

> The more we switch from holding people accountable for their acts to relying on data-driven interventions to reduce risk in society, the more we devalue the ideal of individual responsibility. The predictive state is the nanny state, and then some. Denying people's responsibility for their actions destroys their fundamental freedom to choose their behavior. (p. 177)

What does the future of violence prediction entail? One can only try and predict it.

Chapter 3

It's All About the Base Rates

"The chief reason for our ignorance of the base rates is nothing more subtle than our failure to compute them"

— Paul E. Meehl and Albert Rosen

When prediction using a diagnostic test outperforms simple prediction using base rates, the test is said to be "clinically efficient." This term was first defined by Meehl and Rosen (1955); its importance is often dismissed by proponents of actuarial devices for predicting violent and dangerous behavior. This chapter examines clinical efficiency in such devices. The chapter also provides three equivalent conditions for determining clinical efficiency of a prediction method: (1) Meehl-Rosen (Meehl & Rosen, 1955); (2) Dawes (Dawes, 1962); and (3) the Bokhari-Hubert condition, introduced here.

3.1 Clinical Efficiency

Base rates play an important role in prediction and decision making (Bar-Hillel, 1980; Faust & Nurcombe, 1989; Kahneman & Tversky, 1973; N. Schwarz, Strack, Hilton, & Naderer, 1991). The phrase "clinical efficiency" refers to prediction by a diagnostic test being better than prediction using just base rates (Meehl & Rosen, 1955). If $P(A) \leq 1/2$, then prediction by base rates would be to say consistently that a person does not have "it" because the probability of a correct prediction is $P(\bar{A}) \geq 1/2$ (i.e., the prediction is correct at least half the time). Similarly, if $P(A) > 1/2$, prediction by base rates would be to always

say that the person has "it." Prediction according to the diagnostic test is to say that the person has "it" when the test is positive, and to say the person does not have "it" when the test is negative. To measure how "good" a diagnostic test is, consider the *accuracy* (or *hit rate*) of the test defined as

$$\begin{aligned} P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) &= \left(\frac{n_{BA}}{n_A}\right)\left(\frac{n_A}{n}\right) + \left(\frac{n_{\bar{B}\bar{A}}}{n_{\bar{A}}}\right)\left(\frac{n_{\bar{A}}}{n}\right) \\ &= \frac{n_{BA} + n_{\bar{B}\bar{A}}}{n}. \end{aligned}$$

This expression is just the sum of the main diagonal frequencies from a 2×2 contingency table (e.g., see Table 1.1) divided by the total number of subjects, n.

Assuming $P(A) \leq 1/2$, a general condition can be given for when prediction by a test will be better than prediction by base rates:

$$P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) > P(\bar{A}); \tag{3.1}$$

in words, when the base rate is at most $1/2$, the test should be used for prediction only if the accuracy of the test is greater than the proportion of the population not having "it."

Using the general condition presented in Equation (3.1), there are three important (and equivalent) conditions that can be derived for clinical efficiency. All three conditions involve an attempt to predict an event having a low base rate by using a test possessing less than ideal sensitivity and specificity values; they characterize the circumstances when more accurate prediction would just be to use the larger base rate (i.e., to say the person does not have "it") rather than to rely on the diagnostic test. These three equivalent conditions for base-rate prediction being superior to prediction from the test are attributed to Meehl and Rosen, Dawes, and Bokhari and Hubert; the introduction of this latter condition is the major motivation for the current chapter.

3.1.1 Meehl-Rosen Condition

Assume $P(A) \leq 1/2$. The Meehl-Rosen condition (Meehl & Rosen, 1955) states that it is best to use the test over base rates if and only if

$$P(A) > \frac{1 - P(\bar{B}|\bar{A})}{P(B|A) + (1 - P(\bar{B}|\bar{A}))}, \tag{3.2}$$

or in terms of specificity and sensitivity,

$$\frac{1 - \text{specificity}}{\text{sensitivity} + (1 - \text{specificity})}.$$

Because $1 - P(\bar{B}|\bar{A}) = P(B|\bar{A})$ this condition implies that the test should be used for prediction over base rates if and only if the base-rate probability is larger than the ratio of the false positive rate to the sum of the true positive and false positive rates. The proof of the Meehl-Rosen condition and all other proofs can be found in Chapter A.

If $P(A) > 1/2$, the Meehl-Rosen condition becomes

$$P(\bar{A}) > \frac{1 - P(B|A)}{P(\bar{B}|\bar{A}) + (1 - P(B|A))} = \frac{1 - \text{sensitivity}}{\text{specificity} + (1 - \text{sensitivity})},$$

and the proof is similar.

3.1.2 Dawes Condition

Assume $P(A) \leq 1/2$. The Dawes condition (Dawes, 1962) states that it is best to use the test over base rates if and only if

$$P(\bar{A}|B) < \frac{1}{2}. \tag{3.3}$$

Equivalently, the Dawes condition can be written as $P(A|B) > 1/2$, implying that prediction by the test is better than prediction by base rates if and only if the positive predictive value

is greater than ½. If the positive predictive value is less than ½ (i.e., the Dawes condition fails to hold and it is better to just use base rates for prediction rather than the test), it is more likely that a person does not have "it" than they do even if the test says the person has "it." In other words, given a positive test result there is a higher probability that the person does not have "it" than they do. This has been called the "false positive paradox."

If $P(A) > ½$, then the Dawes condition becomes

$$P(\bar{A}|\bar{B}) > \frac{1}{2};$$

in words, when $P(A) > ½$, the negative predictive value must be greater than ½ for prediction by the test to outperform prediction by base rates. The proof is similar.

3.1.3 Bokhari-Hubert Condition

Assume $P(A) \leq ½$. We will show that it is better to use the test over base rates if and only if differential prediction holds between the row entries in the contingency table: $n_{BA} > n_{B\bar{A}}$ and $n_{\bar{B}\bar{A}} > n_{\bar{B}A}$. In words, when the number of true positives (n_{BA}) is greater than the number of false positives ($n_{B\bar{A}}$) and the number of true negatives ($n_{\bar{B}\bar{A}}$) is greater than the number of false negatives ($n_{\bar{B}A}$), prediction using the test is better than prediction by base rates.

This condition requires no probability calculations and can be seen directly in the contingency table—for base rates to be worse than the test, differential prediction must exist. All three conditions are equivalent; if the Bokhari-Hubert condition holds then the positive predictive value is greater than ½ (due to the Dawes condition). In addition, the Bokhari-Hubert condition implies the negative predictive value is greater than ½; thus, the Bokhari-Hubert condition is equivalent to both the positive and negative predictive values being greater than ½. Unlike the Meehl-Rosen and Dawes conditions, when $P(A) > ½$ the Bokhari-Hubert condition is exactly the same. Thus, if prediction by the test is better than

prediction by base rates, the Bokhari-Hubert condition holds for any $P(A)$.

Relationship to measures of association.

The Goodman-Kruskal lambda coefficient (Goodman & Kruskal, 1954) is a proportional-reduction-in-error measure for predicting a column event (A or $\bar{A}$) from knowledge of a row event (B or $\bar{B}$) over a naïve prediction based solely on marginal column frequencies (n_A and $n_{\bar{A}}$). Thus, the Goodman-Kruskal lambda coefficient can be considered a measure of association between the diagnostic test result and the state of nature. For the 2×2 contingency table (e.g., Table 1.1), lambda is defined as:

$$\lambda_{\text{column|row}} = \frac{\max\{n_{BA}, n_{B\bar{A}}\} + \max\{n_{\bar{B}A}, n_{\bar{B}\bar{A}}\} - \max\{n_A, n_{\bar{A}}\}}{n - \max\{n_A, n_{\bar{A}}\}}.$$

If $\lambda_{\text{column|row}}$ is zero, the maximum of the column marginal frequencies is the same as the sum of the maximum frequencies within rows; therefore, no differential prediction of a column event is made based on knowledge of what particular row an object belongs to. A non-zero $\lambda_{\text{column|row}}$ is an alternative way of specifying the Bokhari-Hubert differential prediction condition. If $\lambda_{\text{column|row}} = 0$, $\max\{n_{BA}, n_{B\bar{A}}\} + \max\{n_{\bar{B}A}, n_{\bar{B}\bar{A}}\} = \max\{n_A, n_{\bar{A}}\}$. If $\max\{n_A, n_{\bar{A}}\} = n_A = n_{BA} + n_{\bar{B}A}$, $n_{BA} \geq n_{B\bar{A}}$ and $n_{\bar{B}A} \geq n_{\bar{B}\bar{A}}$, and the condition fails to hold. Similarly, if $\max\{n_A, n_{\bar{A}}\} = n_{\bar{A}} = n_{\bar{B}\bar{A}} + n_{B\bar{A}}$, $n_{B\bar{A}} \geq n_{BA}$ and $n_{\bar{B}\bar{A}} \geq n_{B\bar{A}}$, and again the condition fails to hold.

An alternative and more popular test of association is based on Pearson's chi-squared statistic (Pearson, 1900b). Although this test can be used for significance testing in a 2×2 contingency table, it says nothing about differential prediction. For instance, this test may show a significant relation between the state of nature (A and $\bar{A}$) and the diagnostic test results (B and $\bar{B}$), but when $\lambda_{\text{column|row}}$ is zero, there is no differential prediction and the use of base rates will outperform the diagnostic test.

Relationship to odds ratio and relative risk.

Odds ratios, or relative odds, are another way of measuring association. The odds of an event is defined as the ratio of the probability that a person has "it" to the probability that a person does not have "it," given a specific diagnostic test result:

$$O_B = \frac{P(A|B)}{P(\bar{A}|B)} = \frac{{}^{n_{BA}}/_{n_B}}{{}^{n_{B\bar{A}}}/_{n_B}} = \frac{n_{BA}}{n_{B\bar{A}}}$$

and

$$O_{\bar{B}} = \frac{P(A|\bar{B})}{P(\bar{A}|\bar{B})} = \frac{{}^{n_{\bar{B}A}}/_{n_{\bar{B}}}}{{}^{n_{\bar{B}\bar{A}}}/_{n_{\bar{B}}}} = \frac{n_{\bar{B}A}}{n_{\bar{B}\bar{A}}}.$$

The first term, O_B, gives the odds of a person having "it" when they tested positive for having "it"; the second term, $O_{\bar{B}}$, gives the odds that a person has "it" when they did not test positive for "it." In Bayesian terms, the odds can be thought of as posterior odds, given the test result; the prior odds is ${}^{P(A)}/_{P(\bar{A})}$. The odds ratio is simply the ratio of the two odds, $OR = {}^{O_B}/_{O_{\bar{B}}}$. Thus, the odds ratio compares which group (B versus $\bar{B}$) is more likely to have "it." If the Bokhari-Hubert condition holds, then $n_{BA} > n_{B\bar{A}} \Leftrightarrow O_B > 1$ and $n_{\bar{B}\bar{A}} > n_{\bar{B}A} \Leftrightarrow O_{\bar{B}} < 1$. This means that if the person tests positive for "it," the odds are greater than not that they do have "it"; if the person tests negative for "it," the odds are greater that they do not have "it" than they do. If $O_B > 1$ and $O_{\bar{B}} < 1$, then $OR > 1$. Therefore, the Bokhari-Hubert condition implies that the odds ratio is greater than one; thus, the odds someone has "it" is greater in the group that tests positive for "it." Of course, none of the entries in the denominators can be zero, but when the Bokhari-Hubert condition holds, only $n_{B\bar{A}}$ has any possibility of being equal to 0.

Relative risk is the ratio of the probability that a person has "it" given they tested positive for having "it" to the probability that a person has "it" given that they did not test

positive for "it." The relative risk is defined as

$$RR = \frac{P(A|B)}{P(A|\bar{B})} = \frac{{}^{n_{BA}}/_{n_B}}{{}^{n_{\bar{B}A}}/_{n_{\bar{B}}}} = \frac{n_{BA}n_{\bar{B}}}{n_{\bar{B}A}n_B}.$$

This ratio is greater than 1 if and only if $n_{BA}n_{\bar{B}} > n_{\bar{B}A}n_B$. If the Bokhari-Hubert condition holds, $n_{BA} > n_{B\bar{A}}$ and $n_{\bar{B}\bar{A}} > n_{\bar{B}A}$.

For this implication to work, $n_{\bar{B}A} > 0$. In summary, if $n_{\bar{B}A}, n_{B\bar{A}} > 0$, the Bokhari-Hubert condition holds if and only if $O_B > 1$ and $O_{\bar{B}} < 1$; the Bokhari-Hubert condition also implies $OR > 1$ and $RR > 1$.

Relationship to diagnostic likelihood ratios.

The positive diagnostic likelihood ratios can be used to assess the performance of a diagnostic test. A positive diagnostic likelihood ratio, DLR_B, provides the likelihood that a positive test (indicating that a person has "it") occurs in an individual who truly does have "it" than one who does not. Similarly, a negative diagnostic likelihood ratio, $DLR_{\bar{B}}$, indicates the likelihood that a negative test (indicating a person does not have "it") occurs in an individual who truly does have "it" than one who does not. The diagnostic likelihood ratios are defined as follows:

$$DLR_B = \frac{P(B|A)}{P(B|\bar{A})} = \frac{{}^{n_{BA}}/_{n_A}}{{}^{n_{B\bar{A}}}/_{n_{\bar{A}}}} = \frac{n_{BA}}{n_{B\bar{A}}}\left(\frac{n_{\bar{A}}}{n_A}\right),$$
$$DLR_{\bar{B}} = \frac{P(\bar{B}|A)}{P(\bar{B}|\bar{A})} = \frac{{}^{n_{\bar{B}A}}/_{n_A}}{{}^{n_{\bar{B}\bar{A}}}/_{n_{\bar{A}}}} = \frac{n_{\bar{B}A}}{n_{\bar{B}\bar{A}}}\left(\frac{n_{\bar{A}}}{n_A}\right).$$

Ideally, a diagnostic test has $DLR_B > 1$ and $DLR_{\bar{B}} < 1$. If the Bokhari-Hubert condition holds, then we know $n_{BA} > n_{B\bar{A}}$; $DLR_B > 1$ if ${}^{n_{BA}}/_{n_{B\bar{A}}} > {}^{n_A}/_{n_{\bar{A}}}$, which is always true if $P(A) \leq {}^1/_2$. Similarly, if the Bokhari-Hubert condition holds, $n_{BA} > n_{B\bar{A}}$ and $DLR_{\bar{B}} < 1$ if ${}^{n_{\bar{B}A}}/_{n_{\bar{B}\bar{A}}} < {}^{n_A}/_{n_{\bar{A}}}$, which is always true if $P(A) \geq {}^1/_2$. Thus, if the Bokhari-Hubert condition holds, at least one of the two ideal diagnostic likelihood ratio conditions also holds.

And what about when the Bokhari-Hubert condition is not met? If the condition fails, it is either because (a) $n_{BA} \leq n_{B\bar{A}}$, or (b) $n_{\bar{B}\bar{A}} \leq n_{\bar{B}A}$, or both (a) and (b) are true. If (a) is true and $P(A) \geq 1/2$, then $DLR_B \leq 1$. Similarly, if (b) is true and $P(A) < 1/2$, then $DLR_{\bar{B}} \geq 1$. The other two situations (where (a) is true but $P(A) < 1/2$ or (b) is true but $P(A) > 1/2$) will depend on the data.

3.2 Predicting Violence and Dangerousness

Predicting violence and dangerousness continues to be a heavily debated topic; the importance of base rates in predicting violent behavior has been discussed elsewhere (Doren, 1998; Vrieze & Grove, 2008; Wollert, 2006), although not all agree (e.g., see G. T. Harris & Rice, 2007). We begin with a numerical example of predicting violence from Monahan et al. (2005) using an actuarial model developed in the MacArthur Violence Risk Assessment Study (Monahan et al., 2001) to attempt validation of the model. This article reports a cross-validated instrument for the diagnostic assessment of violence risk (event B, risk present; event $\bar{B}$, risk absent) in relation to the occurrence of followup violence (event A, violence present; event $\bar{A}$, violence absent) among persons with mental disorders. Table 3.1 displays their results in the form of a 2×2 contingency table.

		State of Nature		
		A (Violence Present)	$\bar{A}$ (Violence Absent)	Totals
Prediction	B (Risk Present)	19	36	55
	$\bar{B}$ (Risk Absent)	9	93	102
	Totals	28	129	157

Table 3.1: 2×2 contingency table for predicting risk of violence among persons with mental disorders (Monahan et al., 2005).

The base rate for violence in this sample is $28/157 = .18 < 1/2$. The authors correctly predicted violence in approximately one-third of their patients ($19/55 = .35$); the authors also correctly predicted nonviolence in about ten of every eleven patients ($93/102 = .91$). Overall,

the authors correctly diagnosed three of every four patients; thus, the accuracy of their test was $^{(19+93)}/_{157} = .71$. Because $P(A) \leq {}^1/_2$, prediction by base rates would be to say that all patients will not commit violence. In doing so, one would be correct 82% of the time for this sample ($P(\bar{A}) = {}^{129}/_{157} = .82$). Because prediction using base rates is better than prediction using their test, all three of the conditions fail, as demonstrated below.

The specificity is $^{93}/_{129} = .72$, the sensitivity of the test is $^{19}/_{28} = .68$, and the base rate (for violence) is $P(A) = .18$. Attempting to verify the Meehl-Rosen condition, we see

$$.18 = P(A) \not> \frac{1 - \text{specificity}}{\text{sensitivity} + (1 - \text{specificity})} = \frac{1 - .72}{.68 + (1 - .72)} = .29,$$

so the condition fails to hold. The positive predictive value (PPV) of the test is PPV $= {}^{19}/_{55} = .35 < {}^1/_2$, so the Dawes condition fails to hold. Finally, because $n_{BA} = 19 < 36 = n_{B\bar{A}}$, the Bokhari-Hubert condition also fails. Another easy way to detect the failure of this latter condition is to note there is no differential prediction because the row entries in the contingency table (Table 3.1) are ordered in the same direction.

The authors also provide "revised estimates" of their sample, reclassifying participants from both groups by "using a slightly more inclusive operational definition of violence" (Monahan et al., 2005, p. 814). (Interestingly, this revised definition differs from the original MacArthur definition of violence used to develop the instrument being validated.) In their revised estimate their accuracy is better (it is exactly equal to $P(\bar{A})$), but all three conditions again fail to hold.

Using the original MacArthur dataset (Monahan et al., 2001), several diagnostic tests were examined for predicting dangerousness. The first is based on arrest history (B: dangerous—one or more prior arrests; $\bar{B}$: not dangerous—no prior arrests). These data are presented in Table 3.2. Three of every four predictions of dangerousness are wrong ($^{294}/_{397} = .74$), and one of every ten predictions of not being dangerous are wrong ($^{39}/_{393} =$

.10); the accuracy of the test is $^{(103+354)}/_{790} = .58$. Using base-rate prediction (i.e., predicting that no one is dangerous) yields an accurate decision for four out of every five predictions ($^{648}/_{790} = .82$). Thus, prediction by base rates is superior to prediction using prior arrest history. This same conclusion is evidenced by noting that differential prediction fails because $n_{B\bar{A}} = 294 > 103 = n_{BA}$ implying that the Bokhari-Hubert condition fails to hold.

		State of Nature		
		A (Dangerous)	$\bar{A}$ (Not Dangerous)	Totals
Prediction	B (Dangerous)	103	294	397
	$\bar{B}$ (Not Dangerous)	39	354	393
	Totals	142	648	790

Table 3.2: 2×2 contingency table for predicting dangerousness based on history of prior violence (Monahan et al., 2001).

A second diagnostic test from the MacArthur data used prior violence. If an individual exhibited prior violence, they were at risk to commit a violent act in the future (B); if the individual did not have any prior history of violence, they were not at risk ($\bar{B}$). The data are shown in Table 3.3. Here, seven out of ten predictions of dangerousness are wrong ($^{106}/_{154} = .69$), and one out of six predictions of being not dangerous are incorrect ($^{128}/_{785} = .16$). The accuracy of this test is $^{(48+657)}/_{939} = .75$; the accuracy of prediction by base rates is $^{763}/_{939} = .81$. Again, this test fails to outperform base-rate prediction, and there is a failure of the Bokhari-Hubert condition: $n_{BA} < n_{B\bar{A}}$.

		State of Nature		
		A (Dangerous)	$\bar{A}$ (Not Dangerous)	Totals
Prediction	B (Dangerous)	48	106	154
	$\bar{B}$ (Not Dangerous)	128	657	785
	Totals	176	763	939

Table 3.3: 2×2 contingency table for predicting dangerousness based on prior arrest history (Monahan et al., 2001).

Kozol et al. published a paper in 1972 entitled *The Diagnosis and Treatment of Dangerousness*. It presented a ten-year study involving nearly 600 male convicted offenders,

most of whom were sex offenders. The authors advised the courts on 592 patients and 435 were released. Based on their clinical assessment and prior history, the authors recommended the release of 386 and opposed the release of 49. In other words, the authors labeled 386 of the patients as not dangerous ($\bar{B}$) and 49 of the patients as dangerous (B). After release, the patients were followed for an average of 43 months. Of the 386 patients recommended for release, 31 committed serious assaultive crimes (A), and 355 did not ($\bar{A}$). For the patients not recommended to be released, but who were nevertheless released, 17 of the 49 committed serious assaultive crimes and 32 did not. The base rate of dangerous was $P(A) = {}^{48}/_{435} = .11$. A 2×2 contingency table summarizing these frequencies is given in Table 3.4.

		State of Nature		
		A (Dangerous)	$\bar{A}$ (Not Dangerous)	Totals
Prediction	B (Dangerous)	17	32	49
	$\bar{B}$ (Not Dangerous)	31	355	386
	Totals	48	387	435

Table 3.4: 2×2 contingency table for predicting dangerousness among criminal patients (Kozol et al., 1972).

As is clear from Table 3.4, the Bokhari-Hubert condition fails to hold; thus, prediction by base rates (i.e., nobody is dangerous) is better than prediction from the test. It is noted that we are not the first to point out flaws of these data; in 1973, Monahan stated that Kozol et al.'s conclusion of reliably diagnosing dangerousness "is, at best, misleading and is largely refuted by their own data" (p. 418). One might make this same unfortunate conclusion for the instruments and variables studied under the MacArthur Violence Risk Assessment Study.

3.3 Conclusion

For more than fifty years conditions have been available to assess when a diagnostic test will outperform base-rate prediction. Meehl and Rosen's (1955) article has been cited

numerous times but the main point appears to be lost on most authors conducting research on violence prediction. The typical way to describe the accuracy of a diagnostic test is by the area under the receiver operator characteristic (ROC) curve; unfortunately, this can be extremely misleading. The implications of an inability to predict violence have meaningful consequences for the individuals involved. Incarcerating an individual based on the results of a diagnostic test that fails to outperform base rates is unethical; as is profiteering from these diagnostic tests.

The Bokhari-Hubert condition presented in this chapter provides a simple condition for determining whether prediction from a diagnostic test outperforms prediction by base rates; this condition is equivalent to those presented by Meehl and Rosen (1955) and Dawes (1962). Unlike these latter two conditions, the Bokhari-Hubert pattern is unchanged with respect to base-rate probability. The Bokhari-Hubert condition also has several relationships with measures of association, such as Goodman-Kruskal's lambda and odds ratios.

The simplicity of the Bokhari-Hubert condition relies on the use of 2×2 contingency tables; this is turn leads to questioning why researchers typically fail to present data in this simple form. Besides its use in assessing the Bokhari-Hubert condition, 2×2 contingency tables provide all the information needed to determine the quality of a diagnostic test, including the area under the ROC curve for a single cutscore.

This chapter also demonstrated that several measures for predicting violence fail to satisfy the BH condition. Unfortunately this appears to be the norm. In a meta-analysis of seventy-three samples, Fazel et al. (2012) determined that the median positive predictive value among the measures examined was .41 suggesting that the instruments failed to satisfy the BH condition in over half of the studies. This is a disconcerting figure that deserves more attention than given, possibly due to the sole reliance on the area under the receiver operating characteristic (ROC) curve (AUC) for measuring test accuracy. The AUC is the topic of the next chapter.

Chapter 4

Hiding Behind the AUC

"Some individuals use statistics as a drunk man uses a lamppost—for support rather than for illumination"

—Andrew Lang

The most common measure for accuracy in methods for predicting violent and dangerous behavior is the receiver operating characteristic (ROC) curve (AUC) and it is often touted for being independent of base rates. This chapter argues that this notion is why they should not be used and are in fact misleading, particularly when the base rate is low. We suggest the use of positive and negative predictive values as a supplement, or even substitute, for the AUC measures.

4.1 Introduction

The receiver operating characteristic (ROC) curve used to evaluate binary classifiers is a plot of the true positive rate (sensitivity) against the false positive rate ($1 -$ specificity). A common measure of how a test performs in practice is the area under the ROC curve (AUC); the AUC represents the probability that a randomly chosen person who has "it" has a higher score than a person who does not (Hanley & McNeil, 1982). We argue that whenever the base rates for a condition being assessed are relatively low, the AUC is inappropriate as a measure for evaluating the adequacy of the actual predictions made from a diagnostic test. This inappropriateness results from the AUC's failure to incorporate information about base rates. The AUC only evaluates the test itself and not how the test actually performs when

used on a specific population having unequal base rates for the presence or absence of the particular condition being assessed. Despite this current widespread adoption, the AUC as a measure of diagnostic adequacy can be misleading when evaluating conditions with differing base rates. These interpretative problems are further compounded when AUC measures are the basic data subjected to a meta-analysis.

In contrast to some incorrect understandings in the literature about the invariance of specificity and sensitivity across samples, it has been known for some time that these two basic measures are subject to a variety of biases (Begg, 1987). For example, sizable subgroup variation can be present in the sensitivity and specificity values for a diagnostic test (this has been called "spectrum bias"; Ransohoff & Feinstein, 1978). In addition, because sensitivity and specificity are calculated from frequencies present in a 2×2 contingency table, it is always best to remember the operation of Berkson's fallacy (Berkson, 1946): the relationship that may be present between two dichotomous variables in one population may change dramatically for a selected sample based on some other variable or condition (e.g., hospitalization, being a volunteer, age, and so on). In short, because ROC measures are generally *not* invariant across groups, however formed, we do not agree with the sentiment expressed in the otherwise excellent review article by Swets et al. (2000b). We quote:

> These two probabilities [sensitivity and specificity] are independent of the prior probabilities (by virtue of using the priors in the denominators of their defining ratios). The significance of this fact is that ROC measures do not depend on the proportions of positive and negative instances in any test sample, and hence, generalize across samples made up of different proportions. All other existing measures of accuracy vary with the test sample's proportions and are specific to the proportions of the sample from which they are taken. (p. 26)

Our general suggestion is to use the positive predictive value (PPV) and the negative predictive value (NPV) to evaluate diagnostic test performance. These two measures incorporate both specificity and sensitivity along with the base rates in the sample for the

presence or absence of the condition under study. In considering the use of positive and negative predictive values, it is important to note that one cannot directly estimate these probabilities is case-control studies as they are predetermined.

4.2 The Area Under the ROC Curve

Although the most commonly used measure of diagnostic adequacy is the area under the ROC curve, our contention, as noted in the Introduction, is that the AUC is not a good measure of clinical efficiency precisely because it does not incorporate base rates. (Clinical efficiency — according to Meehl and Rosen (1955) — refers to prediction by a diagnostic test being better than prediction just using base rates). The AUC is a function of the test itself and not of its use on groups of individuals. To show the "independence" of base rates for the AUC, consider an ROC curve (as in Figure 4.1) defined by a single coordinate pair of sensitivity and (1 − specificity) values. This simple situation might conform to making decisions with a diagnostic test having only a single cutscore (threshold), and where the test is considered "positive" when some score exceeds a particular cutscore, and "negative" otherwise. As shown below, the AUC in this case of one cutscore is just the average of the sensitivity and specificity values, and neither is a function of base rates.

We can see explicitly how different normalizations (using base rates) are used to calculate the AUC and accuracy (or *hit rate*) measures. Letting sensitivity $= P(B|A) = {}^{n_{BA}}/_{n_A}$ and specificity $= P(\bar{B}|\bar{A}) = {}^{n_{\bar{B}\bar{A}}}/_{n_{\bar{A}}}$, the AUC is $\left({}^{P(B|A)+P(\bar{B}|\bar{A})}\right)/2 = \left({}^{n_{BA}/n_A+n_{\bar{B}\bar{A}}/n_{\bar{A}}}\right)/2$; accuracy is $P(B|A)P(A)+P(\bar{B}|\bar{A})P(\bar{A}) = {}^{(n_{BA}+n_{\bar{B}\bar{A}})}/_{n}$. Note that when $n_A = n_{\bar{A}}$ (i.e., when the base rates are equal), the accuracy and AUC measures are identical.

4.2.1 The Wilcoxon Statistic

Another way to interpret what the AUC measures is to note its equivalency to the nonparametric Wilcoxon signed rank-test statistic (Hanley & McNeil, 1982) that compares

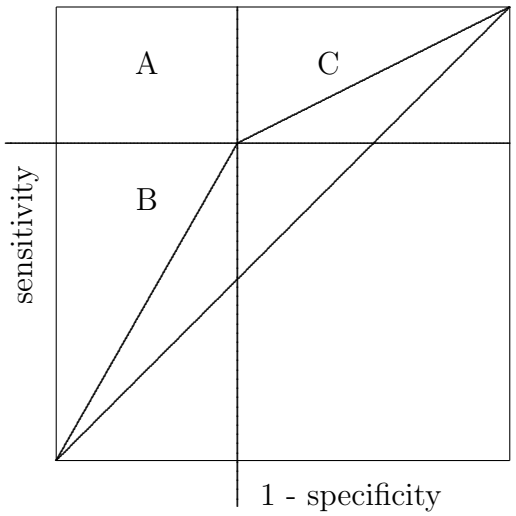

Figure 4.1: An illustrative example of computing the AUC given one cutscore.

randomly selected pairs of observations between two (independent) groups. As an example, suppose we wish to know the probability in a randomly selected pair of people, where one person committed violence and the other did not, that the psychopathy score for the person committing violence is greater than or equal to that for the person not committing violence. For a variable with two ordinal levels, a typical 2×2 contingency table is presented in Table 4.1. The Wilcoxon statistic is

$$\frac{n_{12}n_{21} + \frac{1}{2}\left(n_{22}n_{21}\right) + \frac{1}{2}\left(n_{11}n_{12}\right)}{N_1 N_2}. \tag{4.1}$$

The average of the sensitivity and specificity (i.e., the AUC) is

$$\frac{1}{2}\left(\frac{n_{11}}{N_1} + \frac{n_{22}}{N_2}\right),$$

which, after some algebra, is equal to the Wilcoxon statistic in Equation (4.1).

Similarly, the equivalence between the AUC and the Wilcoxon's statistic can be extended. For example, suppose we have a variable with four ordinal levels, such as for different cutscores on the Screening Version of the Psychopathy Checklist – Revised (PCL:SV; S. D. Hart et al., 1995). A generic table for such a variable is presented in Table 4.2. The

Level	Group 1	Group 2
I	n_{11}	n_{12}
II	n_{21}	n_{22}
Column Totals	N_1	N_2

Table 4.1: A variable with two ordinal levels for two groups.

Wilcoxon statistic for this four-level ordinal variable is

$$\frac{n_{12}\left(n_{21}+n_{31}+n_{41}\right)+\frac{1}{2}\left(n_{11}n_{12}\right)+n_{22}\left(n_{31}+n_{41}\right)+\frac{1}{2}\left(n_{21}n_{22}\right)+n_{32}\left(n_{41}\right)+\frac{1}{2}\left(n_{31}n_{32}\right)+\frac{1}{2}\left(n_{41}n_{42}\right)}{N_1N_2}. \tag{4.2}$$

The AUC can be calculated using the so-called trapezoidal rule, designed to approximate a definite integral. Given a finite number of points (i.e., cutscores) the approximation is exact when all points of the ROC curve are used. Working with the PCL:SV example above, a given cutscore, say c_i, $i = 1, 2, 3, 4$, is represented by the point $(\text{Sens}_{c_i}, 1 - \text{Spec}_{c_i})$ on the ROC plot, where Sens_{c_i} and Spec_{c_i} are the sensitivity and specificity of the test using the cutscore c_i. Beginning from $(0, 0)$ (i.e., predicting that nobody will be violent), the area under the curve from $(0, 0)$ to $(\text{Sens}_{c_4}, 1 - \text{Spec}_{c_4})$ can be determined using the trapezoidal rule:

$$\frac{1}{2}((1-\text{Spec}_{c_4})-0)\left(\text{Sens}_{c_4}+0\right)=\frac{1}{2}\left(1-\frac{n_{32}+n_{22}+n_{12}}{N_2}\right)\left(\frac{n_{41}}{N_1}\right)=\frac{\frac{1}{2}(n_{42}n_{41})}{N_1N_2}.$$

Continuing with the remaining cutscores, it can be shown that this equals the Wilcoxon statistic of Equation (4.2).

Level	Group 1	Group 2
I	n_{11}	n_{12}
II	n_{21}	n_{22}
III	n_{31}	n_{32}
IV	n_{41}	n_{42}
Column Totals	N_1	N_2

Table 4.2: A variable with four ordinal levels for two groups.

4.2.2 Berkson's Bias: An Illustrative Example

In 1978, R. S. Roberts et al. presented one of the first real data examples evidencing Berkson's bias. In reviewing different biases in analytic research, Sackett (1979) presented a table adapted from R. S. Roberts et al. (1978) that showed the relationship between allergic and metabolic disease with fatigue both in the general population and for a selected population hospitalized within the previous six months. These data are given in Table 4.3. The author points out the differences in relative risk (1.89 in the general population versus 0.37 in the hospitalized population), but one might also note that both "tests" fail to meet the Bokhari-Hubert condition (see Chapter 3) that characterizes those situations where it would be more accurate to predict from simple base rates and say that nobody suffered from fatigue than to use allergic and metabolic disease as an indicator of fatigue. That aside, we are presented with a test used in two populations so AUCs can be compared.

The base rates in each population are unequal: $P(A) = .05$ in the general population and $P(A) = .11$ in the hospitalized population. The sensitivities for the two tests are .09 (for the general population) and .04 (for the hospitalized population). The specificities are .95 (for the general population) and .91 (for the hospitalized population). The AUC for the test used in the general population is $^{(.09+.95)}/_{2} = .52$; the AUC for the test used in the hospitalized population is $^{(.03+.91)}/_{2} = .47$. Although neither test is much different from one with no discriminating power (i.e., an AUC = .50), the AUC (.47) for the hospitalized population is actually below the so-called line of discrimination (i.e., the 45° line) in an ROC plot, rendering the test useless (one would be better off reversing the diagnostic test results).

Looking at the negative and positive predictive values it can be seen that the tests fail to outperform base-rate prediction because, for both populations, the PPV is less than $^{1}/_{2}$ (.09 for the general population and .05 for the hospitalized population). The justification for this statement is summarized in the next section.

		General Population		
		Fatigue		
		Yes	No	Row Totals
Allergic and Metabolic	Yes	13	136	149
Disease	No	127	2508	2635
	Column Totals	140	2644	2784

		Hospital Population		
		Fatigue		
		Yes	No	Row Totals
Allergic and Metabolic	Yes	1	21	22
Disease	No	27	208	235
	Column Totals	28	229	257

Table 4.3: An example of Berkson's bias (R. S. Roberts et al., 1978).

4.3 Positive and Negative Predictive Values

The Bokhari-Hubert (BH) condition developed in Chapter 3 states that a diagnostic test will outperform base rates (i.e., $P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) > P(\bar{A})$) if and only if differential prediction holds between the row entries in a 2×2 contingency table (i.e., $n_{BA} > n_{B\bar{A}}$ and $n_{\bar{B}\bar{A}} > n_{\bar{B}A}$). This condition is equivalent to both the positive and negative predictive values being greater than $1/2$; therefore, the negative and positive predictive values presented together determine if the test is outperforming simple base-rate prediction.

The condition where the positive predictive value is greater than one-half, attributed to Robyn Dawes (Dawes, 1962), provides a minimal condition that a diagnostic test must satisfy for it to be better than prediction according to base rates (when $P(A) < 1/2$). Stated in words, assuming the base rate for a person having "it" is less than $1/2$, the positive predictive value must be greater than $1/2$ for prediction by the test to outperform prediction by base rates. If this condition fails to hold, it will be more likely that the person does not have "it" than they do, even when a test result says the person has "it." This has been referred to as the "false positive paradox."

One way to relate the positive predictive value with the sensitivity and the negative predictive value with specificity is by applying Bayes' Theorem. The positive predictive value is

$$P(A|B) = P(B|A)\left(\frac{P(A)}{P(B)}\right),$$

and the negative predictive value is

$$P(\bar{A}|\bar{B}) = P(\bar{B}|\bar{A})\left(\frac{P(\bar{A})}{P(\bar{B})}\right).$$

If $P(A) = P(B)$ and, consequently, $P(\bar{A}) = P(\bar{B})$, the positive predictive value is equal to the sensitivity and the negative predictive value is equal to the specificity. Assuming $P(A) < 1/2$, the general condition for when a diagnostic test outperforms prediction by base rates can be rewritten as

$$P(A|B)P(B) + P(\bar{A}|\bar{B})P(\bar{B}) > P(\bar{B}).$$

The use of these measures will possibly eliminate the terminological confusion about what the term "false positive" might mean. One usual interpretation (that does not take into account the base rates) is (1 − specificity): the probability that the test is positive given that the person does not have "it." Another interpretation (which does account for base rates) is (1 − NPV): the probability that a person has "it" given that the test is negative. Similarly, for a "false negative," the usual interpretation (failing to take base rates into account) is (1 − sensitivity): the probability that the test is negative given that a person has "it"; the other (taking base rates into account) is (1 − PPV): the probability that the person does not have "it" given that the test is positive. By equating $P(A)$ and $P(B)$, the confusions about the meaning of false positive and false negative can be finessed because different interpretations can be given as to what is "false" and what is "positive" and "negative."

When evaluating a diagnostic test where different cutscores can be set, it makes

intuitive sense to set a cutscore so that the proportion of positive decisions is close to the prior probability of a positive decision, $P(A)$, and to then consider the consistency of positive and of negative decisions. The consistency of a positive decision is defined as the proportion of positive decisions that are correct: $P(A \cap B | A \cup B)$; the consistency of a negative decision is the proportion of negative decisions that are correct: $P(\bar{A} \cap \bar{B} | \bar{A} \cup \bar{B})$.

Another possible measure of diagnostic accuracy that does take base rates into account would be an average of the positive and negative predictive values, either weighted or not. Taking the simple average, $(\text{PPV} + \text{NPV})/2$, would correspond to an AUC measure for the single cutscore equalizing the base rates $P(A)$ and $P(B)$. When the BH condition holds, the average of PPV and NPV will be greater than .50, with an upper limit of 1.0. When the BH condition fails, the upper limit of this average is .75 because at least one of PPV or NPV will be at most 1/2. Therefore, any test with an average at or below .50 should not be used because it fails the BH condition. Any test with an average greater than .75 guarantees that the BH condition holds. When the average falls within the range of .50 and .75, the PPV and NPV should be considered separately to determine if the BH condition holds. Reasonable weights for a weighted average of the PPV and NPV would be the base rates of the test so that the weighted average is

$$P(A|B)P(B) + P(\bar{A}|\bar{B})P(\bar{B}) = P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A});$$

this is simply the accuracy of the test. These measures allow comparison between two similar tests using different cutscores.

4.4 AUC Inflation

Consider data from Helmus, Thornton, et al. (2012) using the Static-2002R for estimating five-year risk of recidivism (these data are described in detail in Chapter 7). Table 4.4

presents a 2×2 contingency table when using a moderate-level risk (a score of 5 or higher) as a prediction for recidivism.

		Sexual Recidivism		
		A	$\bar{A}$	Totals
Predicted	B	256	949	1205
Recidivism	$\bar{B}$	85	1319	1404
	Totals	341	2268	2609

Table 4.4: A 2×2 contingency table for predicting sexual recidivism with the Static-2002R. Here, a score of 5 or higher leads to a prediction of recidivism. The results are used to demonstrate Dawes's (1993) properties.

The average of the sensitivity and specificity (i.e., the AUC at this cutscore) is AUC $= {}^{(.75+.58)}/_{2} = .67$; modest, but significant ($p < .001$). The accuracy of predicting recidivism using this cutscore is Acc $= {}^{(256+1319)}/_{2609} = .60 <$ AUC. The positive and negative predictive values of this test are, respectively, PPV $= {}^{256}/_{1205} = .21$ and NPV $= {}^{1319}/_{1404} = .94$; the average of the positive and negative predictive values is ${}^{.21+.94}/_{2} = .58 <$ Acc. Thus, we have the following relationship:

$$\text{AUC} > \text{Acc} > \frac{\text{PPV} + \text{NPV}}{2}. \tag{4.3}$$

4.4.1 Dawes (1993)

Dawes (1993) demonstrates that under certain properties to be discussed shortly, the following inequalities hold:

$$\frac{P(B|A) + P(\bar{B}|\bar{A})}{2} > P(A \cap B) + P(\bar{A} \cap \bar{B}) > \frac{P(A|B) + P(\bar{A}|\bar{B})}{2}. \tag{4.4}$$

Earlier we showed that the left-hand side of Equation (4.4) is the AUC at a given cutscore; the middle term is the accuracy, Acc, and the right-hand side is the average of the positive and negative predictive values. Thus, the equation can be rewritten as Equation (4.3).

As Dawes notes, this suggests that the accuracy observed between a diagnostic test conditioned on the state of nature is reduced when predicting new outcomes from the diagnostic test (i.e., the state of nature conditioned on the diagnostic test). In other words, the average of the sensitivity and specificity will always be larger than the accuracy and this in turn is always greater than the average of the positive and negative predictive values.

Dawes (1993) provides three properties needed to satisfy Equation (4.4); he notes that the second and third property imply the first, but "given the importance of the first ... it is presented separately" (p. 5). The properties are presented in the same order as Dawes (1993).

Property (1)

$$P(B|A) > P(B|\bar{A})$$

The first property states that the sensitivity of the test must be greater than the false positive rate. This is a property that any researcher desires and very few diagnostic tests (in the literature) fail to meet.

Property (2)

$$P(B|A) > P(\bar{B}|\bar{A})$$

The second property states that the sensitivity must be greater than the specificity. This property can be difficult to meet, particularly when the base rate is low. However, it is a reasonable goal for the researcher; paraphrasing Dawes (1993), the researcher is generally interested in predicting the presence of the outcome rather than the absence of it (p. 5).

Property (3)

$$P(A) < \frac{1}{2}$$
$$P(B) < \frac{1}{2}$$

Property (3) states that both the base rate and selection ratio must be less than one-half.

Again, these are reasonable assumptions. Often we are trying to predict the rare event (i.e., $P(A) < 1/2$) so the first part of the property is generally met. The second part of the property certainly should be met especially if $P(A) \ll 1/2$. (See the last section for a discussion on calibration; however, note that if $P(A) = P(B)$ Property (2) cannot be satisfied. In fact, if Properties (2) and (3) are true, $P(B) > P(A)$; this is shown by Dawes and in the proof in Chapter A.)

Given that the three properties are met, then we have the reduction in accuracy given in Equation (4.4); a proof is provided in Chapter A and can also be found in Dawes (1993, Appendix 1). As mentioned, Properties (2) and (3) imply Property (1); this can be noted by the fact that only Properties (2) and (3) are needed to complete the proof. Despite Property (1)'s irrelevance to the proof, out of respect to the wishes of Dawes we keep it. The implications of these properties are important; as Dawes (1993) states

> The properties leading to these results are quite common in psychological investigations and in other social sciences. We are interested in predicting the unusual (base rates $< .50$) from the unusual. When we do so, however, a fundamental asymmetry results. The degree of predictability appears to be systematically greater when the analysis is retrospective than when it is prospective. ... We systematically overestimate. It is that simple. (p. 7)

Consider the hypothetical ROC plot in Figure 4.2. The dotted, gray line is the 45° line called the line of no discrimination. Any point along the ROC curve above (to the left of) this line, satisfies Property (1). The solid black line orthogonal to the line of no discrimination represents the boundary for when Property (2) is met: Any point on the ROC curve above (to the right of) this line satisfies Property (2). The red section of the ROC curve represents the points on the curve satisfying both Properties (1) and (2). If Property (3) is met the cutscores in red represent when the reduction in accuracy is guaranteed to occur.

An interesting phenomenon arises as a consequence of Properties (2) and (3): The negative predictive value will always be larger than the positive predictive value. This is an

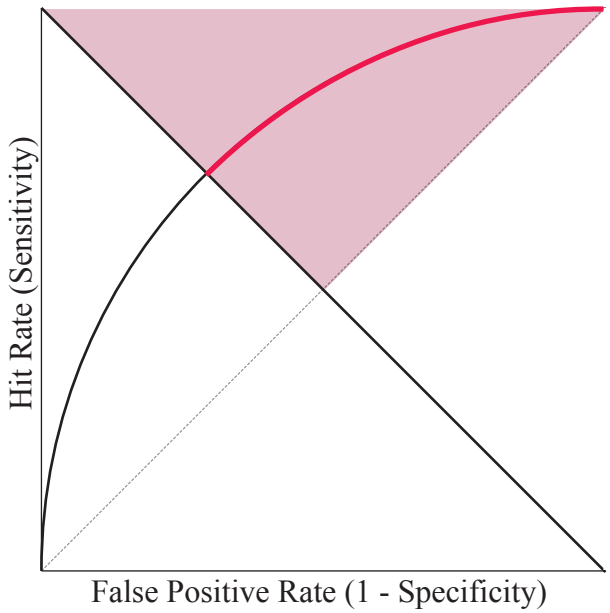

Figure 4.2: A hypothetical ROC plot: the shaded red area represents when Properties (1) and (2) are met; the red line represents the points on the ROC curve satisfying the two properties.

interesting situation because, given Property (2), this seems counterintuitive. To see that this is true refer to the proof provided in Chapter A. The last line in the penultimate set of equations states,

$$\frac{P(\bar{B}|\bar{A})P(\bar{A})}{P(\bar{B})} > \frac{P(B|A)P(A)}{P(B)}.$$

Applying Bayes' Theorem we have

$$P(\bar{A}|\bar{B}) > P(A|B).$$

In other words, if a researcher develops a diagnostic test that has a sensitivity larger than the specificity and the base rate and selection ratio are both less than one-half, then the negative predictive value will be larger than the positive predictive value. In addition there is no restrictive lower bound on the positive predictive value (i.e., we still have PPV > 0).

Returning to the data presented at the beginning of this section (Table 4.4) we first

note that Property (2) is satisfied: $P(B|A) = {}^{256}/_{341} = .75 > P(\bar{B}|\bar{A}) = {}^{1319}/_{2268} = .58$; similarly, Property (3) is satisfied: $P(A) = {}^{341}/_{2609} = .13$ and $P(B) = {}^{1205}/_{2609} = .46$. Also note that as expected, $P(B) > P(A)$ and $P(B|A) > P(B|\bar{A}) = {}^{949}/_{2268} = .42$. Thus, as expected, the inequalities hold as well.

4.5 AUC: Misleading the Way

The following sections provide several examples of the ineffectiveness of the AUC measure in evaluating a test, with particular focus on tests attempting to predict violent or dangerous behavior.

4.5.1 Predicting Violence

Just saying that a measure is "good" because it is independent of base rates does not make it good for the use to which it is being put (or, in computer science jargon, a "bug" does not suddenly become a "feature" by bald-faced assertion). In Chapter 3, an actuarial model of violence risk assessment (Monahan et al., 2005) was used as an example. The 2×2 contingency table is provided in Table 4.5. The average of sensitivity ($P(B|A) = .68$) and specificity ($P(\bar{B}|\bar{A}) = .72$) gives the area under the curve: AUC $= {}^{(.68+.72)}/_{2} = .70$. This number tells us precious little of importance in how the diagnostic test is doing with the cross-validated sample. The positive predictive value is $P(A|B) = .35$ and the negative predictive value is $P(\bar{A}|\bar{B}) = .91$. Because the positive predictive value is less than $1/2$, the test fails the BH condition implying that prediction by base rates ($P(\bar{A}) = .82$) is better than prediction using the test ($P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) = .71$). Based on this test, two out of three predictions of dangerousness are wrong; and one out of eleven predictions of nondangerousness are wrong. It is morally questionable to have one's liberty (or, in some Texas cases, one's life) jeopardized by an assessment of dangerousness that is incorrect two out of three times.

		State of Nature		
		A (Violence Present)	$\bar{A}$ (Violence Absent)	Row Totals
Prediction	B (Risk Present)	19	36	55
	$\bar{B}$ (Risk Absent)	9	93	102
	Column Totals	28	129	157

Table 4.5: A 2×2 contingency table for predicting violence risk among persons with mental disorders (Monahan et al., 2005).

4.5.2 Violence Risk Assessment Study

The data presented in the previous section were used to validate a model developed in the MacArthur Violence Risk Assessment Study (Monahan et al., 2001). This study looked at predicting violence among persons with mental illnesses and had two goals: "to do the best 'science' on violence risk assessment possible and to produce a violence risk assessment 'tool' that clinicians in today's world of managed mental health services could actually use" (Monahan et al., 2001, p. 9). There were several diagnostic measures used in data collection, but the measure that appeared to be the best in terms of predicting violence was the Psychopathy Checklist: Screening Version (PCL:SV). The PCL:SV consists of twelve items; each item is scored 0 (factor not present), 1 (factor may be present or is partially present), or 2 (factor present) during a structured interview. The total score on the PCL:SV ranges from 0 to 24, with higher scores supposedly more predictive of dangerousness and violence. The twelve items on the PCL:SV are identified below using short labels:

1) Superficial

2) Grandiose

3) Deceitful

4) Lacks Remorse

5) Lacks Empathy

6) Doesn't Accept Responsibility

7) Impulsive

8) Poor Behavioral Controls

9) Lacks Goals

10) Irresponsible

11) Adolescent Antisocial Behavior

12) Adult Antisocial Behavior

Table 4.6 displays the results from the MacArthur dataset for each available PCL:SV total score. The base rate for violence in this population is $^{159}/_{860} = .18$; thus, prediction by base rates is to say that no person will commit violence ($P(\bar{A}) = .82$). The table is split into four blocks, representing the three possible cutscores of 6, 12, and 18. Tables 4.7, 4.8, and 4.9 display 2×2 contingency tables using the three different cutscores. Figure 4.3 plots an ROC curve for the three cutscore points.

When using a cutscore of 18, the test does outperform base rates, although minimally. The base rate to three decimals is .815; the accuracy of the test to three decimals is $^{704}/_{860} = .819$. The efficacy of this test is seen in the positive and negative predictive values (.53 and .84, respectively, so the BH condition holds). Note that $P(B) = {}^{55}/_{860} = .06$, which is quite smaller than $P(A)$. The AUC measure tells a different story: the sensitivity is .18, the specificity is .96, and the AUC is .57, the lowest for any of the three tests. One would be hard-pressed trying to equate accuracy with AUC for this test.

4.5.3 Comparing All Cutscores

In the previous subsection, three cutscores were examined in depth but many other cutscores could be considered. When looking at all the cutscores, the positive and negative

PCL-SV Score	Block Yes	Violence at Followup Yes	No	Block No	Row Totals
0		0	34		34
1		1	45		46
2		1	54		55
3		6	48		54
4	18	1	57	328	58
5		4	41		45
6		5	49		54
7		8	51		59
8		10	57		67
9	69	13	38	254	51
10		9	40		49
11		16	31		47
12		13	37		50
13		12	19		31
14		9	14		23
15	43	7	26	93	33
16		3	13		16
17		7	10		17
18		5	11		16
19		10	10		20
20		5	6		11
21	29	4	1	26	5
22		5	5		10
23		0	2		2
24		5	2		7
Column Totals		159	701		860

Table 4.6: PCL:SV data from the MacArthur Risk Assessment Study (Monahan et al., 2001).

predictive values can be useful in determining an "ideal" cutscore for a diagnostic test. Table 4.11 displays the positive and negative predictive values using cutscores ranging from 0 (the minimum PCL:SV total score) to 23 (a cutscore of 24 is equivalent to prediction by base rates; i.e., predicting that no person will commit a violent act); the accuracy and the AUC are also provided.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	141	373	514
	No ($\bar{B}$)	18	328	346
	Column Totals	159	701	860

Table 4.7: Predicting violence using a PCL:SV cutscore of 6.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	72	119	191
	No ($\bar{B}$)	87	582	669
	Column Totals	159	701	860

Table 4.8: Predicting violence using a PCL:SV cutscore of 12.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	29	26	55
	No ($\bar{B}$)	130	675	805
	Column Totals	159	701	860

Table 4.9: Predicting violence using a PCL:SV cutscore of 18.

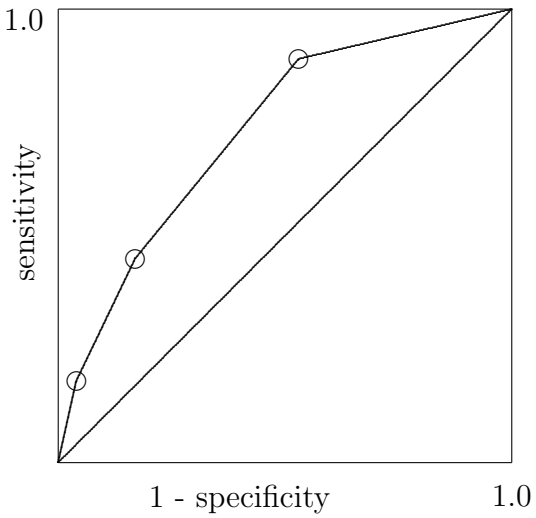

Figure 4.3: Receiver operating characteristic (ROC) curve using three cutscores for PCL:SV in predicting violence.

	Violence	
PCL:SV Score	Yes	No
0–6	18	328
7–12	69	254
13–18	43	93
19–24	29	26
Column Totals	159	701

Table 4.10: Using PCL:SV to predict violence for four cutscores.

All NPV values are greater than .50 for the tests, so cutscores with PPV values greater than .50 meet the BH condition. From Table 4.11, there are six cutscores leading to diagnostic tests that meet the BH condition. Looking at the individual AUC measures, the largest value occurs at a cutscore of 18. Often when determining cutscores, individual AUCs are not used; instead, two common metrics are relied on: the distance from $(\text{Sens}_c, 1-\text{Spec}_c)$ to $(0,1)$: $D = \sqrt{(1-\text{Sens}_c)^2 + (1-\text{Spec}_c)^2}$; or Youden's J (Youden, 1950): $J = \text{Spec}_c + \text{Sens}_c - 1$, where Sens_c and Spec_c are the sensitivity and specificity, respectively, of the test using a cutscore c. Both statistics fall between 0 and 1; minimizing D or maximizing J implies the best cutscore. For the PCL:SV data, the smallest D and the largest J both occur when the cutscore is 8. The accuracy of the test using a cutscore of 8 is .65, and fails to meet the BH condition.

4.5.4 Comparing Tests

We now demonstrate several examples showing the differences in the PPV and NPV despite similar or exactly the same AUC values. First, in their paper discussing the advantages of using ROCs, M. E. Rice and Harris (1995) provided AUC values, as well as several other statistics, for their VRAG instrument after manipulating the base rate. The base rate ranged from .15 to .58. The AUC ranged from .73 to .76, whereas the PPV increased linearly as the base rate increased. The positive predictive value was as low as .36 ($P(A) = .15$), with a maximum of .73 ($P(A) = .58$).

Cutscore	PPV	NPV	Accuracy	AUC
0	0.19	1.00	0.22	0.52
1	0.20	0.99	0.28	0.55
2	0.22	0.99	0.34	0.59
3	0.23	0.96	0.39	0.60
4	0.24	0.96	0.45	0.64
5	0.26	0.96	0.49	0.66
6	0.27	0.95	0.55	0.68
7	0.29	0.94	0.60	0.69
8	0.32	0.92	0.65	0.70
9	0.33	0.91	0.68	0.68
10	0.35	0.90	0.72	0.68
11	0.35	0.88	0.73	0.66
12	0.38	0.87	0.76	0.64
13	0.38	0.86	0.77	0.62
14	0.37	0.85	0.77	0.60
15	0.42	0.85	0.80	0.60
16	0.47	0.85	0.81	0.60
17	0.48	0.84	0.81	0.58
18	0.53	0.84	0.82	0.57
19	0.54	0.83	0.82	0.55
20	0.58	0.83	0.82	0.54
21	0.53	0.82	0.82	0.53
22	0.56	0.82	0.82	0.51
23	0.71	0.82	0.82	0.51

Table 4.11: positive predictive value (PPV), negative predictive value (NPV), their average, accuracy, and AUC for twenty-four cutscores for the PCL:SV.

As a simple demonstration, we show how base rates can dramatically affect the positive predictive value despite the AUC (and sensitivity and specificity at each cutscore) remaining unchanged. We start with a base rate of .50. The results of a hypothetical diagnostic instrument are displayed in Table 4.12. The AUC for this test is .75; the sensitivities at the three cutscores (excluding those predicting nobody and everybody to have "it"; i.e.,

prediction using base rates) are .4, .7, and .9. Similarly, the specificities at the three cutscores are .9, .7, and .4. A plot of the ROC curve can be found in Figure 4.4. All the positive predictive values are greater than or equal to .50 (see Table 4.15).

	Has "It"	
Cutscore	Yes	No
4	200	50
3	150	100
2	100	150
1	50	200
Column Totals	500	500

Table 4.12: Hypothetical diagnostic test results with an AUC of .75; population has a base rate of .50.

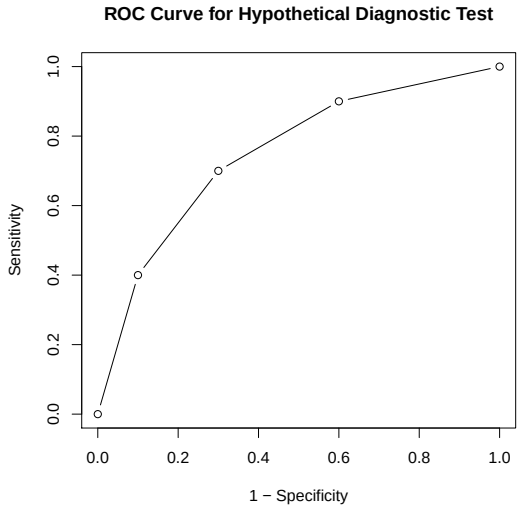

Figure 4.4: ROC curve for ypothetical diagnostic test found in Table 4.12.

We now adjust the base rate of those who have "it" to .40; the results can be found in Table 4.13. The sensitivities and specificities at each cutscore are exactly the same as before; and therefore so is the AUC (.75). The positive predictive values are not the same, however. At each cutscore, the PPV is reduced (see Table 4.15).

Continuing, we adjust the base rates to be .30, .20, .10, .05, and .01. Table 4.14 gives the diagnostic test when the base rate is equal to .01. For each base rate, the hypothetical

	Has "It"	
Cutscore	Yes	No
4	160	60
3	120	120
2	80	180
1	40	240
Column Totals	400	600

Table 4.13: Hypothetical diagnostic test results with an AUC of .75; population has a base rate of .40.

diagnostic test has the same sensitivities and specificities at each cutscore (and same AUC), but, as Table 4.15 summarizes, the positive predictive value is dramatically affected. As the results of this simple exercise demonstrate, simply reporting the AUC can be extremely misleading. In addition, promoting the AUC as a base-rate independent measure without discussing the importance of the positive and negative predictive values leads the reader to ignore the consequences of differing PPVs. If one knows that the test has an AUC of a certain value and the base rate is very low, then the clinician should be aware that the PPV is going to be poor.

	Has "It"	
Cutscore	Yes	No
4	4	99
3	3	198
2	2	297
1	1	396
Column Totals	10	990

Table 4.14: Hypothetical diagnostic test results with an AUC of .75; population has a base rate of .01.

	Base Rate						
Cutscore	.50	.40	.30	.20	.10	.05	.01
4	.80	.73	.63	.50	.31	.17	.04
3	.70	.61	.50	.37	.21	.11	.02
2	.60	.50	.39	.27	.14	.07	.01
1	.50	.40	.30	.20	.10	.05	.01

Table 4.15: Positive predictive values for hypothetical diagnostic tests with differing base rates but fixed sensitivity and specificity (and AUC).

In the last example the sensitivity and specificity were exactly the same at each cutscore and consequently, the AUC was the same. This need not be the case and the final illustrative example demonstrates this (see also Mossman, 2006a). Here, the PCL:SV measure divided into four subgroups from Table 4.10 is used to predict violence; for simplicity we will consider each of the four ranges of PCL:SV scores as individual cutscores so that the test possesses only four cutscore values. Recall the base rate for violence in this sample is $P(A) = {}^{159}/_{860} = .18$. As demonstrated, using a cutscore of 19–24 to predict violence has an accuracy of .82, slightly better than prediction using base rates; thus, the test meets the BH condition when using a cutscore of 19–24. The AUC measure for this test is .7296.

Now suppose the data were slightly different, as presented in Table 4.16. The only differences between the two datasets are the frequencies at each PCL:SV score; the base rate is unchanged (i.e., $P(A) = .18$). In this hypothetical situation, no cutscore meets the BH condition. The best cutscore (in terms of accuracy of the test) is the 13–18 PCL:SV scoring range. Using this cutscore, the accuracy is ${}^{(85+595)}/_{860} = .79 < .82$. The AUC for this test is .7295, nearly the same as for the original test. Both tests differ significantly ($p < .0001$) from one with no discriminating power (AUC = .50), and are not significantly different from each other ($p = .996$); however, only one of the tests has any merit for predicting violence (i.e., the original test).

Now suppose that the definition of violence was redefined to be less stringent and

	Violence	
PCL:SV Score	Yes	No
0–6	19	329
7–12	55	266
13–18	66	28
19–24	19	78
Column Totals	159	701

Table 4.16: Hypothetical PCL:SV distribution. The base rate for violence is the same as in Table 4.10, but the number at each cutscore differs.

we ended up with the hypothetical results in Table 4.17. In this situation we have twice as many "violent" individuals (the base rate for violence is now $P(A) = {}^{318}/_{860} = .37$), but the number at each cutscore is the same as before. In this scenario, the prediction using the test outperforms the base-rate prediction at each cutscore, but the AUC measure is slightly worse than the two other examples (AUC = .7282). The AUC measure significantly differs from a test with no discrimination power ($p < .0001$), but does not differ significantly from the original test ($p = .926$).

	Violence	
PCL:SV Score	Yes	No
0–6	60	286
7–12	124	199
13–18	92	44
19–24	42	13
Column Totals	318	542

Table 4.17: Hypothetical PCL:SV distribution. The number at each cutscore is the same as in Table 4.10, but the base rate is different.

Given these three tests, how is one to choose the "best"? If only the AUC measures are provided, there is no way to distinguish between the three; the reader would be left to conclude that all three tests are good (because they differ significantly from a test with no discrimination power), but that is not the case. Table 4.18 displays the the accuracy and the AUC measure of each test at each cutscore. Scenario 1 consists of the original

data (Table 4.10), scenario 2 refers to the hypothetical example with the same base rate for violence as the original data (Table 4.16), and scenario 3 refers to the hypothetical example with a less stringent definition of violence (Table 4.17). From Table 4.18, the three tests can be compared at each cutscore, or each test across the three cutscores. Again, for each test at each cutscore the AUC does not provide much information regarding test accuracy. For instance, in scenario 1 as the cutscore increases, the accuracy increases but the AUC decreases. For scenario 2, the AUC based on a cutscore of 0–6 is larger than the AUC using a cutscore of 19–24 (.68 vs. .50), but the accuracy provides an opposite conclusion (.55 vs. .75). We again conclude that the AUC is not a good indication of test accuracy.

Scenario	Accuracy	AUC
	Cutscore: 0–6	
Scenario 1	.55	.68
Scenario 2	.55	.68
Scenario 3	.63	.67
	Cutscore: 13–18	
Scenario 1	.76	.64
Scenario 2	.79	.69
Scenario 3	.72	.66
	Cutscore: 19–24	
Scenario 1	.82	.57
Scenario 2	.75	.50
Scenario 3	.66	.55

Table 4.18: Accuracy and AUC for three scenarios of violence prediction at three different cutscores.

4.6 Calibration

Calibration means that the selection ratio is equal to the base rate; that is, $P(A) = P(B)$ ($\Leftrightarrow P(\bar{A}) = P(\bar{B})$). Assuming $P(A) \leq 1/2$, because $P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) = P(A|B)P(B) + P(\bar{A}|\bar{B})P(\bar{B})$, the general condition for when prediction by a diagnostic test

is better than prediction by base rates can be rewritten as

$$P(A|B)P(B) + P(\bar{A}|\bar{B})P(\bar{B}) > P(\bar{B}).$$

Given this scenario, the general condition above is equivalent to differential prediction between the *column* entries: $n_{BA} > n_{\bar{B}A}$ and $n_{\bar{B}\bar{A}} > n_{B\bar{A}}$ (see Appendix A). Thus, when prediction by a diagnostic test is superior to prediction by base rates and $P(A) = P(B)$, differential prediction holds between both the row entries and the column entries.

One way to determine an ideal cutscore is to consider when $P(B) \approx P(A)$ (Cook, 2007). When a diagnostic measure is calibrated, two measures one might consider are the consistency of positive decisions (i.e., the proportion of correct positive decisions among all positive decisions: $P(A \cap B|A \cup B)$) and the consistency of negative decisions (i.e., the proportion of correct negative decisions among all negative decisions: $P(\bar{A} \cap \bar{B}|\bar{A} \cup \bar{B})$). If the BH condition holds, then $n_{BA} > n_{B\bar{A}} = n_{\bar{B}A}$ and $n_{\bar{B}\bar{A}} > n_{\bar{B}A} = n_{B\bar{A}}$ (i.e., differential prediction holds between both the rows and columns). In this situation, it is easy to show that both consistency measures must be at least 1/3 (see Appendix A).

In Chapter 3, the implications of the BH on diagnostic likelihood ratios was discussed. If, in addition to the BH condition being met, $P(A) = P(B)$, then it can be shown that both the positive and negative diagnostic likelihood ratios are ideally met (i.e., $DLR_B > 1$ and $DLR_{\bar{B}}$), regardless of $P(A)$ (see Appendix A). In Chapter 3 it was also shown that when a diagnostic test meets the BH condition at a given cutscore, both the positive and negative predictive values are greater than 1/2. When $P(A) = P(B)$ as well, both the sensitivity and specificity are also greater than 1/2, implying that the AUC is greater than 1/2.

4.6.1 Example

As a numerical example of when $P(A) \approx P(B)$, consider the PCL:SV with a cutscore of 13. The 2×2 contingency table is shown in Table 4.19. The base rate for nonviolence is

$P(\bar{A}) = {}^{701}/_{860} = .815$ and the probability of a nonviolent prediction is $P(\bar{B}) = {}^{700}/_{860} = .814$. The consistency of positive decisions for the test is $P(A \cap B | A \cup B) = {}^{60}/_{(60+100+99)} = .23$; the consistency of negative decisions is $P(\bar{A} \cap \bar{B} | \bar{A} \cup \bar{B}) = {}^{601}/_{(601+100+99)} = .75$. Three out of four negative decisions are consistent, but only about one of four positive decisions is consistent.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	60	100	160
	No ($\bar{B}$)	99	601	700
	Column Totals	159	701	860

Table 4.19: Predicting violence using a PCL:SV cutscore of 13.

Some might argue (e.g., G. T. Harris & Rice, 2013) that the base rates do not matter and that the test is independent of them, so then why calibrate? But the test should be used on a similar population which displays similar charateristics and similar base rates; if not, then the generalizability of the test is worth questioning.

4.7 Conclusion

The AUC measure is a poor indicator of a diagnostic test's accuracy. When used in populations with differing base rates, and in particular in populations with low base rates, the AUC measure can be very misleading. When only the AUC is presented, the reader is forced to base an assessment of the test with a statistic that does a poor job of judging. If AUCs are to be presented, they should be accompanied by the positive and negative predictive values. The PPV and NPV allows a determination of whether the test outperforms base-rate prediction, something that is not possible with the AUC. The AUC also does a poor job of determining an optimal cutscore for a given diagnostic test. Other statistics, such as the distance measure or Youden's J also fail to determine an optimal cutscore in terms of test accuracy. In contrast, the average of PPV and NPV is a better

indicator of how well a test does at a given cutscore. Others have also advocated the use of the PPV and NPV(e.g., see McCusker, 2007); an alternative presentation of the positive and negative predictive values are documented in Frederick and Bowden (2009).

The AUC is often touted as being independent of base rates, but there is evidence that the sensitivity and specificity do vary, and often quite dramatically, across different base rates and that the positive and negative predictive value do not vary as much as might be expected (Brenner & Gefeller, 1997; Leeflang, Bossuyt, & Irwig, 2009). The "independence of base rates" argument is one the most heavily cited reasons to use the AUC, but this argument may be as misleading as the AUC itself. Other issues associated with the AUC have also been noted (Guggenmoos-Holzmann & van Houwelingen, 2000; Hand & Anagnostopoulos, 2013; Marzban, 2004; Vecchio, 1966; Vrieze & Grove, 2008).

In short, it is misleading to use the AUC as an indicator of a good test; the AUC can disguise a poor test as a good one, with obvious consequences for individuals diagnosed by the test. Employing a test with a "significant" AUC measure that fails to outperform base-rate prediction to assess whether an individual will be violent or dangerous is ethically questionable. Predicting violence is a difficult and important task and methods for prediction should be continually improved, but settling for diagnostic tests whose predictions fail to outperform base-rate predictions is unacceptable.

Chapter 5

Lack of Cross-Validation

"What we've got here is ... failure to cross-validate"

Cross-validation is an important evaluation strategy for predictive modeling; without it, a predictive model is likely to be overly optimistic. As Meehl and Rosen (1955) state, "If a psychometric instrument is applied solely to the criterion groups from which it was developed, its reported validity and efficiency are likely to be spuriously high" (p. 194). This chapter discusses different cross-validation methods. To demonstrate the importance of cross-validation, several predictive models are constructed with data from the MacArthur Violence Risk Assessment Study (Monahan et al., 2001) and compared with the original (non-cross-validated) Classification of Violence Risk assessment tool. The results show that the predictive models' measures of accuracy (AUC, misclassification error, sensitivity, specificity, positive and negative predictive values) worsen when applied to a testing sample than compared with the training sample used to fit the model. In addition, unless false negatives (i.e., incorrectly predicting an individual to be nonviolent) are considered more costly than false positives (i.e., incorrectly predicting an individual to be violent), the models generally make few predictions of violence. The implications of these results are discussed; skepticism regarding non-cross-validated results is encouraged.

We begin by introducing several widely-used methods for cross-validation. Data are presented from the MacArthur Violence Risk Assessment Study (VRAS) that were used to

develop the Classification of Violence Risk (COVR) assessment tool. Based on this dataset we construct a main effects logistic regression, a linear and a quadratic discriminant analysis, and several classification tree models and demonstrate the process of cross-validation for each.

5.1 An Introduction to Cross-Validation

Cross-validation is an important tool for prediction. It allows the researcher to estimate a prediction tool's accuracy in practice. Assessing the accuracy of a model with the same data used to create the model will give biased—and overly optimistic—estimates of accuracy; cross-validation is a strategy to mitigate such bias.

Assume we have a dataset $(\mathbf{X}, \mathbf{y})$, where $\mathbf{X}$ is an $n \times p$ matrix containing n observations measured across p predictor variables, and $\mathbf{y}$ is an $n \times 1$ vector containing n observations measured on a single outcome variable (e.g., the outcome of violence). In this scenario the outcome variable is known; this is typically referred to as *supervised learning* (in contrast, when $\mathbf{y}$ is unknown we have *unsupervised learning*). In violence prediction the methods commonly fall into the supervised learning category—the outcome for the subject (whether he or she committed an act of violence) is known when the model is being constructed.

In prediction, interests are in modeling $\mathbf{y}$ as a function of $\mathbf{X}$; it is assumed that for some function f,

$$\mathbf{y} = f(\mathbf{X}) + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ represents the vector of random error terms; it has mean 0 and is uncorrelated with the set of predictor variables. The primary goal is to estimate $f(\mathbf{X})$ so

$$\hat{\mathbf{y}} = \hat{f}(\mathbf{X}).$$

The total error in prediction, $\mathbf{y} - \hat{\mathbf{y}}$, can be divided into two types of components: *reducible*

error ($f(\mathbf{X}) - \hat{f}(\mathbf{X})$) and *irreducible error* ($\boldsymbol{\varepsilon}$). Based on mean-squared error,

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{y} - \hat{\mathbf{y}})^2\right] &= \mathbb{E}\left[(f(\mathbf{X}) + \boldsymbol{\varepsilon} - \hat{f}(\mathbf{X}))^2\right] \\
&= \mathbb{E}\left[(f(\mathbf{X}) - \hat{f}(\mathbf{X}))^2\right] + 2\mathbb{E}\left[(f(\mathbf{X}) - \hat{f}(\mathbf{X}))\boldsymbol{\varepsilon}\right] + \mathbb{E}(\boldsymbol{\varepsilon}^2) \\
&= \underbrace{\mathbb{E}\left[(f(\mathbf{X}) - \hat{f}(\mathbf{X}))^2\right]}_{\text{reducible error}} \quad + \quad \underbrace{\mathbb{V}(\boldsymbol{\varepsilon})}_{\text{irreducible error}},
\end{aligned}$$

where $\mathbb{E}(\cdot)$ and $\mathbb{V}(\cdot)$ represent the expected value and variance, respectively. These two types of error determine the accuracy of predictions. Although the term $\mathbb{V}(\boldsymbol{\varepsilon})$ is unknown and cannot be reduced (hence the term "irreducible error"), the reducible error can be minimized; this is the goal of prediction. If the predicted function perfectly matches the true function, the total mean squared error is equal to $\mathbb{V}(\boldsymbol{\varepsilon})$; thus, $\mathbb{V}(\boldsymbol{\varepsilon})$ represents a lower bound for the total error. In practice $f(\mathbf{X})$ is not known, so one can only hope to get close to this lower bound.

In constructing a predictive model, there are several available measures (e.g., the mean squared error, the coefficient of determination [R^2], the proportion of predictions correct) for assessing how well the model fits the data. It is important to note that this error is often associated with the sample that is relied on to construct the model and not on predictions in an independent sample. If the model is constructed for purposes of prediction, what is most relevant is the model's predictive accuracy on new data. Suppose there is a given measure of accuracy, say γ, for assessing the model and this measure was obtained with the same data used to construct the model. One way of evaluating a model's predictive ability is to gather new data and measure how accurate the predictions are; that is, a new accuracy measure γ' is obtained. The difference between γ and γ' represents the drop in how well the model predicts (assuming a larger γ is associated with better accuracy; typically, $\gamma - \gamma' > 0$); this drop is known as *shrinkage*. Rather than assessing predictive accuracy with the same data relied on to build the model, the original data can be randomly split into

two parts: the *training data* and the *testing data*. The training data is for constructing the model; the testing data is for estimating the predictive accuracy of the model. This process is typically more efficient in terms of time and cost than collecting new data after the model is developed.

5.1.1 Cross-Validation Methods

Several methods are available for cross-validation. The data can be split so that a specified proportion, say q, of the data is present in the training set. Thus, qn observations are in the training data and $(1-q)n$ are in the testing data. This can be carried out multiple times choosing a different training and testing set each time. Because each replication of this process will produce results that vary, it is common to compute an average across all replications.

K-fold Cross Validation.

K-fold cross validation involves splitting the data into K subsets; the training set consists of the union of $K-1$ subsets, and the testing set is defined by the remaining observations. This process is repeated so that each subset acts once as the testing sample. The simplest form of K-fold cross validation is to let $K = 2$: the training set contains half of the observations, and the testing set the other. The most computationally costly form is to let $K = n$, so that each observation acts as the testing sample; this is commonly known as *leave-one-out cross-validation* (LOOCV). A disadvantage of K-fold validation is that the variance of the estimate can be relatively large compared to other estimates; it is, however, approximately unbiased (see Hastie, Tibshirani, & Friedman, 2009, p. 242).

Mote Carlo Validation.

Monte Carlo validation (or *repeated random sub-sampling validation*) randomly splits the data into a training set and a testing set. The model is fit with the training data and

accuracy assessed with the testing data. This is repeated for a different random split over a specified number of replications. The predictive accuracy of the model is determined by averaging across all replications. One advantage of Monte Carlo validation is that the number of training and testing sets is not dependent on the number of replications. In K-fold cross-validation, the size of each dataset is predetermined by K. The obvious disadvantage to Monte Carlo validation is that there is no guarantee each observation will be a part of the training and testing sets.

5.1.2 Resampling Methods

Bootstrapping.

The *bootstrap* (Efron, 1979) is a resampling technique for estimating the sampling distribution for some statistic or for an estimator of some parameter of interest. The process involves randomly sampling a specified number of times from the original data with replacement. For instance, when the dataset contains n observations, one can randomly resample n' observations (typically $n' = n$) from the dataset with replacement. This implies that a given observation could be sampled more than once, and likely will be. An estimate of the parameter is made from the sample; this process is repeated, say B times. For example, when the parameter of interest is γ, each of the $b = 1, \ldots, B$ replications of bootstrap sampling process will produce an estimate of γ, namely $\hat{\gamma}^{(b)}$. The bootstrap method allows construction of the standard error (and thus a confidence interval) for the population parameter of interest. Thus, one obtains a better sense of the variability present in an accuracy estimate, $\hat{\gamma}$.

Jackknife.

The *jackknife* procedure (Tukey, 1958) like the bootstrap, is a resampling method applied to reduce bias and estimate the variance of parameter estimates. The jackknife procedure splits the observations into K groups—much like in K-fold cross-validation—and

the parameter is estimated removing one of the K groups; this is carried out for each group and a new estimate is made defined as the difference between the parameter estimate with all observations and the one leaving out the kth group. For example, if $\hat{\gamma}$ is the parameter estimate for the entire sample and $\hat{\gamma}_{-k}$ is the estimate with the kth group removed, the new estimate is $\hat{\gamma}_k^* = K\hat{\gamma} - (K-1)\hat{\gamma}_{-k}$. These estimates can be used to construct a standard error (and again a confidence interval) for the population parameter, γ. This procedure also reduces bias in an initially biased estimator.

5.2 The MacArthur Violence Risk Assessment Study

As discussed in Chapter 1, the Classification of Violence Risk (COVR) assessment tool is based on an iterative classification tree (ICT) model. Relying on the proportion of violent individuals at each leaf to estimate the probability of committing an act of violence, the authors classified subjects as low-risk when their estimated probability of violence was less than half the base rate (.09) and high-risk if the estimated probability was more than twice the base rate (.37). After the authors developed their first classification tree, 477 patients were classified as either high- or low-risk; the patients not fitting into either of these two groups were selected to construct a second classification tree model. This second tree classified 119 of the previously unclassified subjects into high- or low-risk groups. The third iteration classified 63 more individuals and the fourth, 60 more. The fifth and final tree failed to classify any additional subjects into the two categories leaving 719 individuals classified and 220 unclassified. This ICT process constructed nine more times and the five best models—as chosen from a logistic regression model—-classified individuals into five groups: very high risk, high risk, average risk, low risk, and very low risk with expected violence probabilities of .76, .56, .26, .08, and .01, respectively.

The authors did not cross-validate their model. As Monahan et al. (2001) state on page 106, "Dividing the sample leaves fewer cases for the purpose of model construction"

and, quoting Gardner et al. (1996b), "wastes information that ought to be used estimating the model." When their ICT models were constructed in the late 1990s and early 2000s, computing power was not what it is now, but LOOCV on a dataset of 939 was certainly possible (although possibly not in the version of SPSS they relied on). With today's computing power there is no reason not to cross-validate one's model or to argue that LOOCV "wastes" data. As will be shown, cross-validated error can be drastically different than the misclassification error for the initially constructed model (i.e., the resubstitution error).

As noted, Monahan et al. (2001) cite a source when they make their disingenuous remark, so their reasoning did not necessarily originate with them. Looking at the Gardner et al. (1996b) article referenced in the quote above, a footnote on page 43 states that a bootstrap cross-validation was performed on the authors' logistic regression model, a perfectly reasonable alternative. Monahan et al. also performed a bootstrap analysis to estimate the variability of the predictions; 1,000 bootstrap samples helped estimate 95% confidence intervals for the probability-of-violence point estimates given above.

5.2.1 VRAS Data

Based on data discussed in Monahan et al. (2001), several predictive models are constructed and used to demonstrate the process of cross-validation. Data from the MacArthur Violence Risk Assessment Study (VRAS) were initially discussed in Chapter 1; here, we represent it with more detail. The data are observations on 939 patients discharged from inpatient psychiatric facilities based in Pittsburgh, Kansas City, and Worcester, MA. The ages of the patients range from 18–40 (Mean = 29.9; Median = 30.0). Of the 939 patients, 538 (57%) were male; 645 (69%) were White, 273 (29%) were African-American, and 21 (2%) were Hispanic.

The response variable (`Violence`) is a binary outcome variable representing whether an act of violence took place within the follow up period (`Violence` = 1 if an act of vi-

olence occurred; `Violence` = 0 if not)[1]. Thirty-one predictor variables were included based on the results from the main effects logistic regression and iterative classification tree models in Monahan et al. (2001). The variables used in Monahan et al.'s (2001) main effects logistic regression model are the Barratt Impulsiveness Scale (BIS) non-planning subscale (`BISnp`); the Brief Psychiatric Rating Scale (BPRS; activation subscale: `BPRSa`; hostile-suspiciousness subscale: `BPRSh`; and total score: `BPRSt`); child abuse seriousness (`ChildAbuse`); employed prior to hospitalization (`Emp`); father's drug use (`DadDrug`); prior arrest history (frequency, `PriorArr`); presence of grandiose delusions (`GranDel`); involuntary hospitalization admission status (`LegalStatus`); proportion of social network members who are also mental health professionals (`snmhp`); Novaco Anger Scale (NAS) Behavioral Subscale (`NASb`); loss of consciousness due to head injury (`Consc`); Psychopathy Checklist: Screening Version (PCL:SV) total score (`PCL`); DSM-III-R checklist: drug abuse (`DrugAbuse`); threat/control override symptoms (`tco`); violent fantasies with escalating seriousness (`FantEsc`); and violent fantasies with single target focus (`FantSing`). The variables used in their initial CART model are self-reported violence two months prior to hospitalization (`RecViol2`); alcohol or drug abuse (`SubAbuse`); admission reason (suicide, `Suicide`); father ever arrested (`DadArr`); any previous head injury (`HeadInj`); violent fantasies with target present (`FantTarg`); diagnosis of schizophrenia (`Schiz`); age of patient (`age`); level of functioning (`Function`); arrested since age 18 for property crime (`PropCrime`); MacArthur Perceived Coercion Scale (PCS, `PCS`); threats at admission (`Threats`); and number of negative relationships (`NegRel`). The authors also constructed a "clinically feasible" model, excluding risk factors that were not readily available or involved lengthy instruments (e.g., `PCL`).

All statistical analyses were carried out in MatLab (MATLAB, 2013) with preprocessing done in `R` (R Core Team, 2012). The data are available for download through the MacArthur Research Network website (`http://www.macarthur.virginia.edu/`

[1] All variable coding that follows is based on our coding (see Appendix C.1).

`risk.html`); the dataset for the present analysis was directly obtained from the MacArthur researchers—it is a "cleaned-up" version from the statistician on the project. Monahan et al. (2001) relied on mean (for continuous variables) and mode (for categorical variables) substitution for handling missing data (see p. 93 for details) in their main effects logistic regression model; thus, the same was done for all variables in this analysis. The best attempt was made for preprocessing the data to match that in Monahan et al.'s analysis. All R and MatLab code, including the preprocessing of the data, can be found in Appendix C.2.

As a way of comparing how close our variables match those of the MacArthur authors, Pearson product-moment correlation coefficients were compared to those found in Chapter 5 of Monahan et al. (2001; see Table 5.2, Table 5.3 and Table 5.5). Table 5.1 displays the estimated correlations for each predictor variable with the response variable as well as the reported correlations from Monahan et al.. Note that all correlations are computed before missing value imputation. Although not a foolproof method for confirming that the variables were preprocessed in the same manner, it certainly does indicate discrepancies that may exist. There is a lot of agreement (at least to two decimal places), but it is not complete. Eight of the 31 correlations disagree, seven to only one-hundredths of a correlation. The largest discrepancy is between prior head injury ($r = .04$ vs. $r' = .06$).

5.3 Main Effects Logistic Regression Model

Monahan et al. (2001) constructed a main effects logistic regression (MELR) model to predict violence that was fit with forward-stepwise variable selection with a $p < .05$-threshold for retaining predictor variables. The present analysis constructs an MELR model as well but one fitted with only the variables from the final model given by Monahan et al. (see the discussion in the previous section). Table 5.2 displays the estimated coefficients (to three decimals for comparative purposes) and the estimated odds ratio for the MELR models for both the present analysis and that in Monahan et al. (2001). In addition, the corresponding

Variable	r	r'
BISnp	.05	.05
BPRSa	−.08	−.08
BPRSh	.08	.08
BPRSt	−.04	−.04
ChildAbuse	.14	.14
Emp	−.05	−.05
DadDrug	.15	.16
PriorArr	.24	.24
GranDel	−.01	−.01
LegalStatus	.11	.11
SNMHP	−.10	−.10
NASb	.17	.16
Consc	.09	.10
PCL	.26	.26
DrugAbuse	.16	.17
tco	−.09	−.10
FantEsc	.13	.13
FantSing	.10	.10
RecViol2	.14	.14
SubAbuse	.18	.18
Suicide	−.01	−.01
DadArr	.14	.15
HeadInj	.04	.06
FantTarg	.12	.12
Schiz	−.12	−.12
Age	−.07	−.07
Function	−.01	−.01
PropCrime	.11	.11
PCS	.03	.03
Threats	.06	.06
NegRel	.05	.06

Table 5.1: Pearson product-moment correlations of predictor variable with response variable, Violence, in reanalyzed dataset (r) and reported correlations in Monahan et al. (2001) (r').

95% confidence intervals are provided for the current analysis.

	Present Study				VRAS	
Variable	$\hat{\beta}$	CI	$\widehat{OR}$	CI	$\hat{\beta}$	$\widehat{OR}$
Intercept	-2.900	$(-4.218, -1.583)$	—	—	-2.814	—
BISnp	-0.028	$(-0.053, -0.003)$	0.97	$(0.95, 0.997)$	-0.031	0.97
BPRSa	-0.151	$(-0.275, -0.027)$	0.86	$(0.76, 0.97)$	-0.164	0.85
BPRSh	0.117	$(0.039, 0.196)$	1.12	$(1.04, 1.22)$	0.127	1.14
BPRSt	-0.033	$(-0.065, -0.002)$	0.97	$(0.94, 0.998)$	-0.033	0.98
ChildAbuse	0.373	$(0.169, 0.578)$	1.45	$(1.18, 1.78)$	0.427	1.53
Emp	-0.477	$(-0.870, -0.085)$	0.62	$(0.42, 0.92)$	-0.530	0.59
DadDrug	0.737	$(0.195, 1.279)$	2.09	$(1.22, 3.59)$	0.779	2.18
PriorArr	0.298	$(0.137, 0.459)$	1.35	$(1.15, 1.58)$	0.286	1.33
GranDel	0.711	$(0.037, 1.385)$	2.04	$(1.04, 4.00)$	0.826	2.28
LegalStatus	0.511	$(0.122, 0.901)$	1.67	$(1.13, 2.46)$	0.500	1.65
SNMHP	-1.856	$(-3.322, -0.390)$	0.16	$(0.04, 0.68)$	-1.704	0.18
NASb	0.038	$(0.008, 0.068)$	1.04	$(1.01, 1.07)$	0.038	1.04
Consc	0.520	$(0.007, 1.032)$	1.68	$(1.01, 2.81)$	0.551	1.73
PCL	0.898	$(0.484, 1.313)$	2.46	$(1.62, 3.72)$	0.876	2.40
DrugAbuse	0.381	$(-0.070, 0.831)$	1.46	$(0.93, 2.30)$	0.449	1.58
tco	-0.900	$(-1.571, -0.229)$	0.41	$(0.21, 0.80)$	-0.412	0.66
FantEsc	0.676	$(0.035, 1.316)$	1.97	$(1.04, 3.73)$	0.648	1.90
FantSing	0.565	$(0.057, 1.073)$	1.76	$(1.06, 2.92)$	0.628	1.87

Table 5.2: Estimated coefficients for the main effects logistic regression model ($\hat{\beta}$) and odds ratios ($\widehat{OR}$) for present analysis (with confidence intervals [CI]) and those presented in Monahan et al. (2001, Table 5.1).

The estimated model for the present analysis differs slightly from that presented by Monahan et al. (2001). All variables in the present analysis were found to be significant ($p < .05$) except DrugAbuse ($p = .11$); this is unlike the results in Monahan et al. (2001) in that all estimated beta coefficients were significant ($p < .05$). Predicted probabilities ranged from .003 to .91; the VRAS predicted probabilities ranged from .002 to .93. The authors used a predicted probability cutscore of half the sample base rate (.09) and twice the sample base rate (.37) to classify low- and high-risk patients, respectively. In doing so,

57.1% (536 individuals) of their sample was classified under the two risk categories. Similarly, the present model classified 55.2% (518) of the patients as high risk or low risk. From these results, the 2×2 contingency Table 5.3 can be constructed; those who fall into the high risk group are predicted to be violent, those who fall in the low risk group are not. The base rate for the subsample is $^{98}/_{518} = .189$; the resubstitution error is less, $^{(67+23)}/_{518} = .174$. This result implies that the model outperforms base rate prediction; that is, the accuracy of the model is better than naïve prediction using only base rate information (i.e., predicting all individuals to be nonviolent; see Chapter 3). Given that the base rate for violence in the sample is less than one-half, a model that outperforms base rate prediction is equivalent to one with a positive predictive value greater than one-half (Dawes, 1962, see also Chapter 3); thus, it is guaranteed that at least half of the predictions of violence are correct. Thus, outperforming base rate prediction should be a minimal requirement. The sensitivity and specificity are, respectively, .765 and .840. The positive and negative predictive values are $^{75}/_{142} = .528$ and $^{353}/_{376} = .939$, respectively. The model, when used on the training data, satisfies the Bokhari-Hubert (BH) condition (see Chapter 3); that is, the model outperforms prediction using base rates. The results of the LOOCV are given in Table 5.4; there were five fewer individuals classified as low risk. The base rate is $^{94}/_{513} = .183$; the LOOCV error is $^{(72+24)}/_{514} = .187$, slightly larger meaning that the cross-validated model no longer meets the BH condition. In addition to the increase in overall error, the validated model also shows decreased sensitivity ($^{70}/_{94} = .745$), specificity ($^{347}/_{419} = .828$), positive predictive value ($^{70}/_{142} = .493$), and negative predictive value ($^{347}/_{371} = .935$). Note that because the test fails to outperform base rate prediction, the positive predictive value is less than one-half implying that an individual predicted to be violent is more likely than not to be nonviolent.

Instead of splitting the data by so-called low- and high-risk individuals, a cutscore of .50 might be used to classify violence, representing a "more likely than not" probability. The resubstitution error is $^{(29+130)}/_{939} = .169$ (Table 5.5); the model outperforms base-rate prediction. The sensitivity is $^{46}/_{176} = .261$; the specificity is $^{734}/_{763} = .962$. The positive and

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	75	67	142
	No ($\bar{B}$)	23	353	376
	Column Totals	98	420	518

Table 5.3: Predicting violence with a main effects logistic regression model. If a patient had a predicted probability greater than two times the base rate (.37), a prediction of violence was made; if the predicted probability was less than half the base rate (.09), a prediction of no violence was made.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	70	72	142
	No ($\bar{B}$)	24	347	371
	Column Totals	94	419	513

Table 5.4: LOOCV results for predicting violence with a main effects logistic regression model. If a patient had a predicted probability greater than two times the base rate (.37), a prediction of violence was made; if the predicted probability was less than half the base rate (.09), a prediction of no violence was made.

negative predictive values are, respectively, $^{46}/_{75} = .613$ and $^{734}/_{864} = .850$. The LOOCV error is .182 (Table 5.6), slightly less than the base rate; thus, the cross-validated model also satisfies the BH condition and outperforms base rate prediction. However, the sensitivity is low indicating that the model fails to correctly identify most (over three quarters) of violent individuals. The cross-validated sensitivity ($^{39}/_{176} = .222$), specificity ($^{734}/_{763} = .955$), positive predictive value ($^{39}/_{73} = .534$), and negative predictive value ($^{729}/_{866} = .842$) are all less. The ROC curve for the main effects model is shown in Figure 5.1. The AUC for the non-cross-validated logistic regression model is .79 (.75, .83); the cross-validated AUC is .77 (.72, .81). Both AUC values are significantly different than .50 and significantly different from each other ($p < .001$). The authors reported an AUC of .81 for their logistic regression model.

Next, the data are randomly split into two parts, the training set (70% of the original sample, or 658 observations) and the testing set (30% of the original sample, or 281 obser-

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	46	29	75
	No ($\bar{B}$)	130	734	864
	Column Totals	176	763	939

Table 5.5: Predicting violence with a main effects logistic regression model. If a patient had a predicted probability greater than .50, a prediction of violence was made; otherwise a prediction of no violence was made.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	39	34	73
	No ($\bar{B}$)	137	729	866
	Column Totals	176	763	939

Table 5.6: LOOCV results for predicting violence with a main effects logistic regression model. If a patient had a predicted probability greater than .50, a prediction of violence was made; otherwise a prediction of no violence was made.

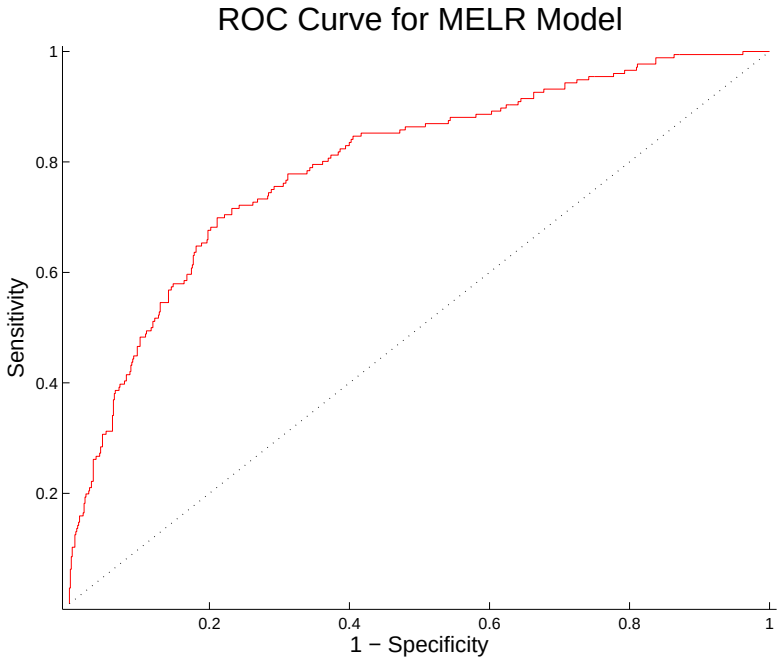

Figure 5.1: ROC plot for the main effects logistic regression model. The AUC is .79.

vations). The base rate for violence in the training set is .188; in the testing set, .185. The model fit to the training data is then used to predict violence in the testing data. These

results (based on a .50 cutscore) are given in Table 5.7, where Table 5.8 displays the results for the high-risk and low-risk cutscores.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	12	17	29
	No ($\bar{B}$)	40	212	252
	Column Totals	52	239	281

Table 5.7: Predicting violence with a main effects logistic regression model on the testing sample. If a patient had a predicted probability greater than .50, a prediction of violence was made; otherwise a prediction of no violence was made.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	22	31	53
	No ($\bar{B}$)	9	101	110
	Column Totals	31	132	163

Table 5.8: Predicting violence with a main effects logistic regression model on the testing sample. If a patient had a predicted probability greater than two times the base rate (.37), a prediction of violence was made; if the predicted probability was less than half the base rate (.09), a prediction of no violence was made.

Setting the cutscore to .50 and fitting the model to the training data, the model's sensitivity and specificity are $^{40}/_{124} = .323$ and $^{518}/_{534} = .970$, respectively; the positive and negative predictive values are $^{40}/_{56} = .714$ and $nf518602 = .860$, respectively; and the resubstitution error is $^{100}/_{658} = .152$, much lower than the base rate. When the model is used to classify training observations as high-risk and low-risk, the sensitivity and specificity are now $^{58}/_{75} = .773$ and $^{260}/_{303} = .858$, respectively; the positive and negative predictive values are $^{58}/_{101} = .574$ and $^{260}/_{277} = .939$, respectively; and the resubstitution error is $^{60}/_{378} = .159$, again less than the base rate for violence. The AUC for the model fit to the training data is .80 $(.75, .84)$. (Note that the training data results are not provided as tables.)

These results suggest the model is doing well at predicting violence; however, it is important to note that they are based on the same data used to fit the model. Fitting

the model to the testing data (with a cutscore of .50), the sensitivity and specificity are $^{12}/_{52} = .231$ and $^{212}/_{239} = .889$, respectively; the positive and negative predictive values are $^{12}/_{29} = .414$ and $^{212}/_{252} = .84$, respectively; and the misclassification error is $^{(17+40)}/_{281} = .203$, slightly larger than the base rate meaning the model fails to meet the BH condition. Classifying testing observations as high-risk and low-risk, the sensitivity and specificity are $^{22}/_{31} = .710$ and $^{101}/_{132} = .765$, respectively; the positive and negative predictive values are $^{22}/_{53} = .415$ and $^{218}/_{255} = .918$, respectively; and the misclassification error is $^{(31+9)}/_{163} = .245$, again implying the model fails to satisfy the BH condition. The AUC is .74 $(.66, .82)$.

From these results we find that the model performs far worse on the testing sample, demonstrating the importance of cross-validation and the inflated accuracy when measured with the same data that created the model; the sensitivity, specificity, positive and negative predictive values, and AUC all decrease. Furthermore, the model no longer meets the BH condition. The logistic regression model does not seem nearly as promising when cross-validated on new data.

5.4 Discriminant Analysis

This section applies discriminant analysis to classify individuals from the VRAS data, for both a linear fit (i.e., assuming the covariances among the two populations—nonviolent and violent—are equal) and a quadratic fit (i.e., the covariances are allowed to be unequal).

For a linear discriminant function on the entire dataset, the AUC is .80 and the resubstitution error is .176, less than the base rate; the LOOCV error is .190, slightly larger than the base rate so the model satisfies the BH condition and outperforms base rate prediction. The data were then split into a training and testing sample. The training sample again contains 70% of the original sample (658 observations); the base rate for violence is $^{124}/_{658} = .188$. The testing sample contains the remaining 30% (281 observations); the base rate for violence is $^{52}/_{281} = .185$. After fitting a linear discriminant model to the training data,

it was used to predict the observations in the testing data; the 2×2 contingency Table 5.9 displays the results with a cutscore of .50. The sensitivity and specificity of the tests are, respectively, $^{14}/_{52} = .269$ and $^{210}/_{229} = .917$; the positive and negative predictive values are, respectively, $^{14}/_{33} = .424$ and $^{210}/_{248} = .847$; the misclassification error is $^{(19+38)}/_{281} = .203$; and the AUC is .74 $(.67, .83)$. The model fails to outperform base prediction and fails to correctly identify a large number of violent individuals.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	14	19	33
	No ($\bar{B}$)	38	210	248
	Column Totals	52	229	281

Table 5.9: Predicting violence with a linear discriminant analysis model on the testing data. If a patient had a predicted probability greater than .50, a prediction of violence was made; otherwise a prediction of no violence was made.

Applying a quadratic discriminant function, the model is again fit to the entire dataset. The AUC is .91 and the resubstitution error is .145, quite an improvement over the linear model. However, the LOOCV error is .227, much higher than the linear model. Based on the same training and testing samples as the linear model, the results of Table 5.10 are obtained. The model performs much worse on the testing data than the linear model. The sensitivity and specificity are, respectively, $^{17}/_{52} = .327$ and $^{202}/_{229} = .882$; the positive and negative predictive powers are, respectively, $^{17}/_{44} = .386$ and $^{202}/_{237} = .852$; the misclassification error is $^{(27+35)}/_{281} = .221$; and the AUC is .69 $(.60, .77)$. Again, the model does not outperform base rate prediction and identifies a only small proportion of violent individuals.

These results again show the differences in resubstitution error when cross-validating. Additionally we see that a more flexible model (i.e., using a quadratic classifier) does better on the training data than the less flexible model (i.e., using a linear classifier), but worse on the testing sample. This is an indication that the more flexible model overfits the data.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	17	27	44
	No ($\bar{B}$)	35	202	237
	Column Totals	52	229	281

Table 5.10: Predicting violence with a quadratic discriminant analysis model on the testing data. If a patient had a predicted probability greater than .50, a prediction of violence was made; otherwise a prediction of no violence was made.

5.5 Classification and Regression Trees

The last type of modeling discussed involves classification and regression tree (CART) analyses as described in Chapter 1. A classification tree is first developed as done in Monahan et al. (2001) before proceeding to more sophisticated methods. For the initial CART analysis, SPSS (IBM Corportation, 2012) was used to implement the Chi-Squared Automatic Interaction Detection (CHAID; Kass, 1980) decision tree; a decision tree is also constructed in MatLab. Subsequent analyses will be with an ensemble learning method known as bagged decision trees. For these latter CART analyses, MatLab (MATLAB, 2013) is used exclusively.

Before beginning, it is necessary to discuss how a classification tree classifies observations and how this process is related to the costs of false positives and negatives.

5.5.1 Misclassification Costs

Suppose in a given terminal node there are n observations, of which n_k are from class k ($k = 1, \ldots, K$). An observation is classified into class k based on the modal class at that given terminal node; thus, if $n_k > n_{k'}$ for all $k \neq k'$, all observations within the terminal node are classified as belonging to class k. The empirical posterior probability for each class can be defined as the number of observations in the terminal node coming from a particular class divided by the total number of observations; thus, the estimated posterior probability

is

$$\hat{P}(k|\mathbf{x}) = \frac{n_k}{n},$$

where $\mathbf{x}$ is a vector of predictor variables associated with the observation. Given this definition, an observation, y, is classified as coming from class k when $\hat{P}(k|\mathbf{x}) > \hat{P}(k'|\mathbf{x})$ for all $k \neq k'$.

As an addition to the classification process, costs can be assigned to misclassifications; the cost function is labeled $C_j(k)$ and represents the cost of classifying an observation into class j when it belongs in class k (note that $C_k(k) = 0$). By including a cost function, an observation is classified into class k by minimizing

$$\sum_{k=1}^{K} \hat{P}(k|\mathbf{x}) C_j(k)$$

across all j. Note that when $C_j(k)$ is the same for all $k = 1, \ldots, K$ (i.e., the costs are equal across all classes), the previous situation obtains and an observation is classified based on the modal class.

Given two classes (i.e., $k = 1, 2$; for example, 1 could represent nonviolent individuals and 2 violent individuals) an observation is classified into class $k = 2$ when

$$\hat{P}(2|\mathbf{x}) C_1(2) > \hat{P}(1|\mathbf{x}) C_2(1).$$

With respect to classification of nonviolent and violent individuals, $C_1(2)$ and $C_2(1)$ are, respectively, the costs associated with a false negative and a false positive. Alternatively, the above inequality can be written as

$$\frac{C_1(2)}{C_2(1)} > \frac{\hat{P}(1|\mathbf{x})}{\hat{P}(2|\mathbf{x})}.$$

The lower bound, $\hat{P}(1|\mathbf{x})/\hat{P}(2|\mathbf{x})$, is the conditional odds in favor of the event 1; for example,

the odds in favor that an individual is not violent, given the data.

If $C_1(2) = C_2(1)$, an observation is classified as coming from class 2 when $\hat{P}(2|\mathbf{x}) > \hat{P}(1|\mathbf{x})$, or equivalently,

$$\frac{\hat{P}(2|\mathbf{x})}{\hat{P}(1|\mathbf{x})} > 1.$$

Bayes Theorem allows this to be rewritten as

$$\frac{\frac{\hat{P}(\mathbf{x}|2)\hat{P}(2)}{\hat{P}(\mathbf{x})}}{\frac{\hat{P}(\mathbf{x}|1)\hat{P}(1)}{\hat{P}(\mathbf{x})}} = \frac{\hat{P}(\mathbf{x}|2)\hat{P}(2)}{\hat{P}(\mathbf{x}|1)\hat{P}(1)} > 1.$$

Considering $\hat{P}(\mathbf{x}|2)$ and $\hat{P}(\mathbf{x}|1)$ fixed, the classification cutscore can be changed by adjusting $\hat{P}(1)$ and $\hat{P}(2)$; these probabilities are the sample base rates (note that for $k = 1, 2$, $\hat{P}(2) = 1 - \hat{P}(1)$). Thus, adjusting the prior probabilities is an equivalent way for adjusting costs.

As noted earlier, Monahan et al. (2001) suggested the cutscore for classification of high-risk individuals be twice the sample base rate of violence (approximately .37). This implies that an individual is classified as violent when the individual belongs to a terminal node where $\hat{P}(2|\mathbf{x}) > .37$. This implicitly assigns unequal costs to false positives and negatives; let $\hat{P}(2|\mathbf{x}) = 2P(A) = .37$ (and consequently, $\hat{P}(1|\mathbf{x}) = 1 - 2P(A) = .63$) so

$$\frac{C_1(2)}{C_2(1)} = \frac{1 - 2P(A)}{2P(A)} = \frac{.63}{.37} = 1.67.$$

By lowering the cutscore to .37 for classification of violence, the authors have implied that false negatives are 1.67 times worse than false positives. By letting the cutscore equal .50 the misclassification error in each terminal node is ensured to be less than .50, implying that the BH condition is satisfied in the decision tree for the training data.

Most authors of actuarial measures are reluctant to discuss the costs of false positives versus false negatives (Mossman, 2006b, 2013; Vrieze & Grove, 2008); an exception is Richard Berk. In his book, *Criminal Justice Forecasts of Risk: A Machine Learning Approach* (Berk, 2012), he suggests that

> *the costs of forecasting errors need to be introduced at the very beginning when the forecasting procedures are being developed* [original emphasis]. Then, those costs can be built into the forecasts themselves. The actual *forecasts* [original emphasis] need to change in response to relative costs. (p. 20)

In the examples Berk provides (regarding parole release), he suggests that the ratio of false negatives to false positives be as high as twenty to one (see also Berk, 2011).

5.5.2 VRAS CART Model in SPSS

Although Monahan et al. (2001) included all 134 potential risk factors in their analysis, for simplicity only the 31 predictor variables previously discussed and provided in detail in Appendix C.1 are included here. As in Monahan et al. (2001), the minimum leaf size was set to 50 with no limit on the tree depth and a significance level of $p < .05$ for variable selection. A Bonferroni adjustment was applied for significance values; it is unknown whether the authors implemented a Type-I-error correction in their analyses. Finally, a K-fold cross-validation is implemented, not done by the authors. The syntax can be found in Appendix C.2.

The classification tree is displayed in Figure 5.2; the results are poor and not similar to what was identified by the authors. With a .50 cutscore (as determined by the proportion of violent individuals in a given terminal node), both the resubstitution and cross-validation errors were .187, or equal to the base rate, because all cases were classified as nonviolent in the decision tree (i.e., no terminal node contained a sample with more than 50% having committed an act of violence). The proportion of violent individuals in each terminal node varied from 0 to .392. Using the previously noted values of .37 and .09 for high-low classifications to predict violence, the contingency Table 5.11 is constructed (see also Figure 5.2 where the low-risk groups are within dashed boxes and the high-risk groups are within bold boxes). Of the 939 patients, 444 were classified as high- or low-risk (compare this to the 518 classified in the MELR model). The CART model performs poorly; only 62.1% of patients

were correctly classified. The base rate for violence in the subset is .236, but the model classifies nearly 60% of the sample as high risk.

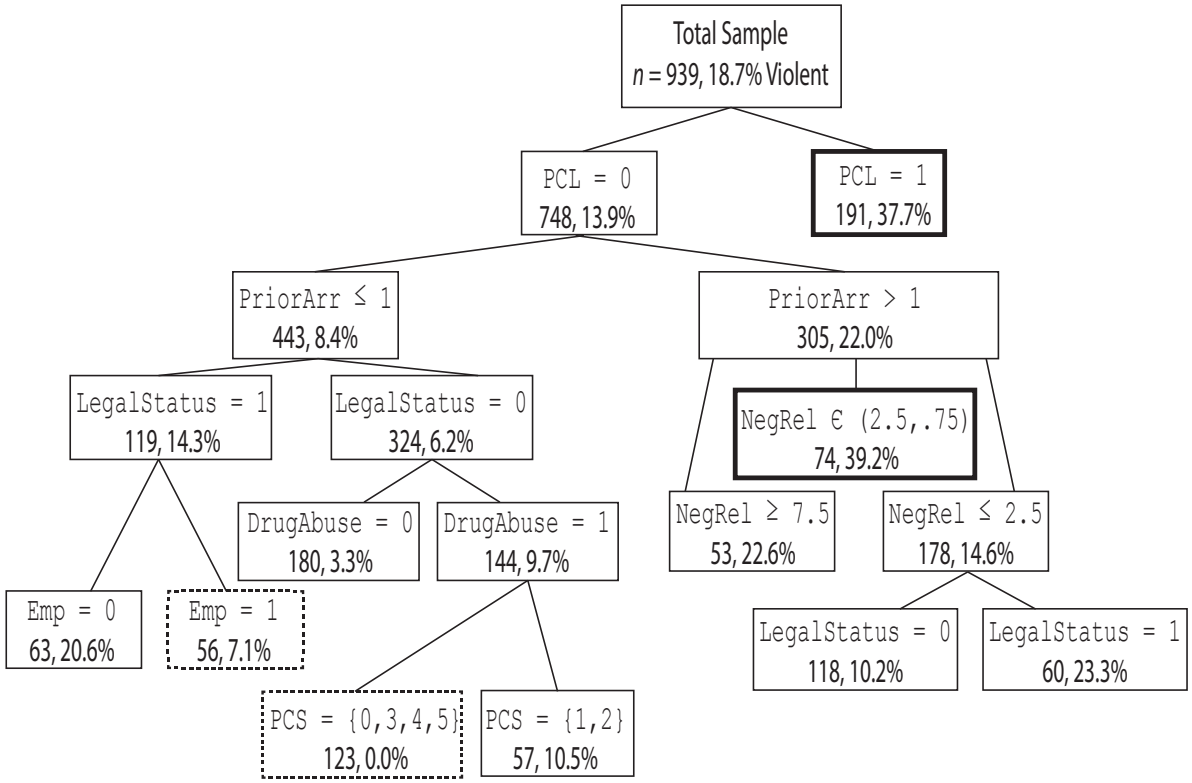

Figure 5.2: CHAID decision tree for MacArthur data (Monahan et al., 2001) using SPSS and following similar guidelines set forth by the authors. The bold boxes represent terminal nodes with a proportion of violent individuals greater than twice the base rate (.37); the dotted boxes represent terminal nodes with a proportion less than half the base rate (.09).

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	101	164	265
	No ($\bar{B}$)	4	175	179
	Column Totals	105	339	444

Table 5.11: Predicting violence with a CHAID decision tree model. If a patient had a predicted probability greater than .50, a prediction of violence was made; otherwise a prediction of no violence was made.

5.5.3 VRAS CART Model in MatLab

We begin by constructing a simple classification tree to determine the minimum leaf size for each node by defining the leaf size over the $\log_{10}$-space, ranging from 10 to 100 at ten intervals; thus each point, $i = 1, \ldots, 10$, is equal to $10^{(8+i)/9}$. The general intent here is to obtain an idea of minimum leaf size based on the data. To quantify this decision, LOOCV error is estimated; the minimum leaf size plotted against the cross-validation error is provided in Figure 5.3 (left plot). Note that after the point $10^{16/9} \approx 59.9$, the cross-validated error is the same as the base rate for violence (labeled with a dashed horizontal line and labeled "BR" on the vertical axis). This implies that the classification tree is not providing any information (i.e., everyone is classified as non-violent), so a minimum leaf size is restricted to be less than 60. Next, each integer between 1 and 60 was assessed for the minimum leaf size. Because K-fold cross-validation randomly selects K subsamples, and the minimum leaf size will be influenced by this randomness, LOOCV is adopted at the cost of increased computing time. The plot displaying the minimum leaf size versus the LOOCV error is given in the right-hand side of Figure 5.3. Several minimum leaf sizes give a CV error less than the base rate; the minimum CV error of .178 is obtained at the two minimum leaf sizes of 41 and 42. A minimum leaf size of 41 produces a decision tree with a minimum leaf of 42 (i.e., there is no tree produced that has a leaf with the specified minimum of 41).

Based on a minimum leaf size of 42, the first classification tree constructed in MatLab is shown in Figure 5.4. The resubstitution error (with a cutscore of .50) is .175, implying the misclassification of 164 patients; the LOOCV-error is a slightly higher .178. Both measures indicate the model is outperforming base rate prediction (the base rate for violence in the sample is .187). The high-risk, low-risk classification for violence prediction discussed earlier is essentially useless with this tree because no terminal node contains a subsample with a rate of violence less than half the base rate. Based on a .50 cutscore, 46 individuals are classified as violent (29 of whom are) and the rest nonviolent.

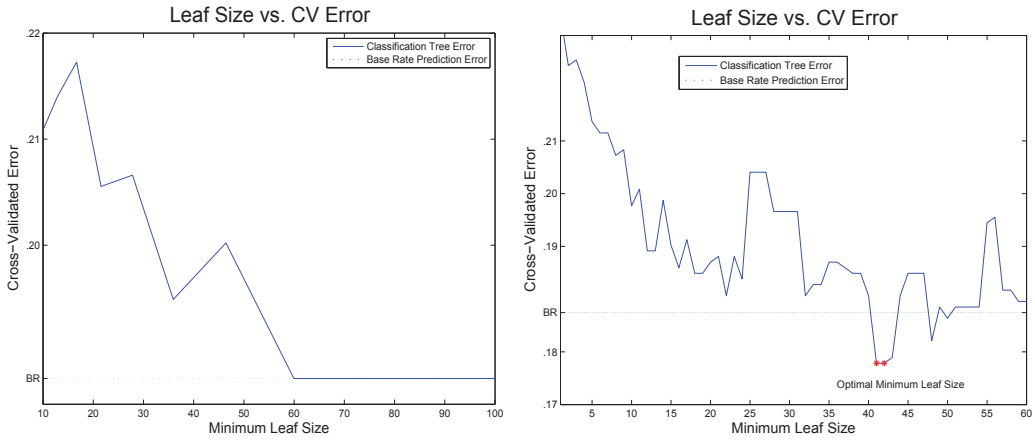

Figure 5.3: Determining the minimum leaf size for a classification tree with K-fold cross-validation error. The figure on the left is over $\log_{10}$-space ($K = 10$). The results helped determine a more restricted range for the results in the right-hand plot; the minimum leaf size is determined to be 42 ($K = n$).

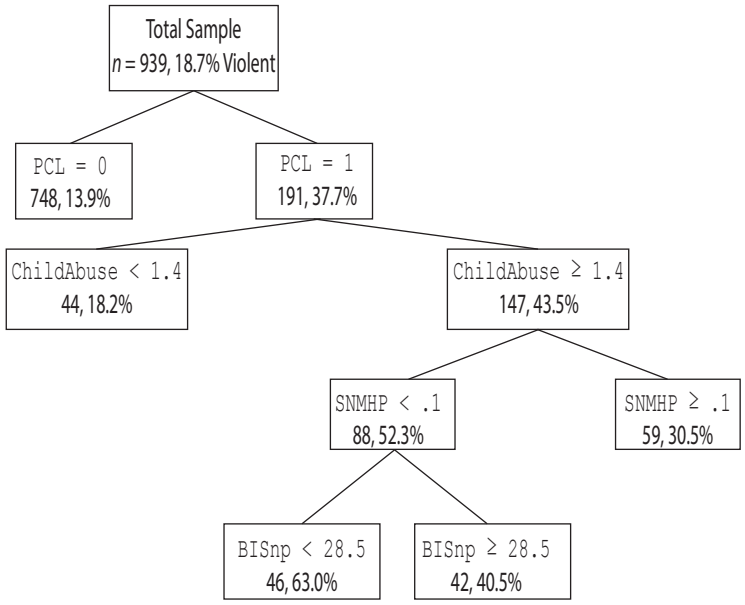

Figure 5.4: Classification tree with an optimal minimum leaf size of 42 and setting equal costs.

The same analyses were repeated but the cost matrix was set so the cutscore for classifying individuals was twice the base rate (i.e., .37); thus, false negatives are considered to be about 1.67 times more costly than false positives (see the earlier section). Through similar analyses presented earlier, the minimum leaf size was determined to be 26; Figure 5.5

displays the tree. The resubstitution error for this model is .177—less than the base rate; the cross-validated error is .257—much larger than the base rate. Unlike before when costs of false positives and negatives were considered equal, a minimum leaf size of 50 does produce a tree (not shown).

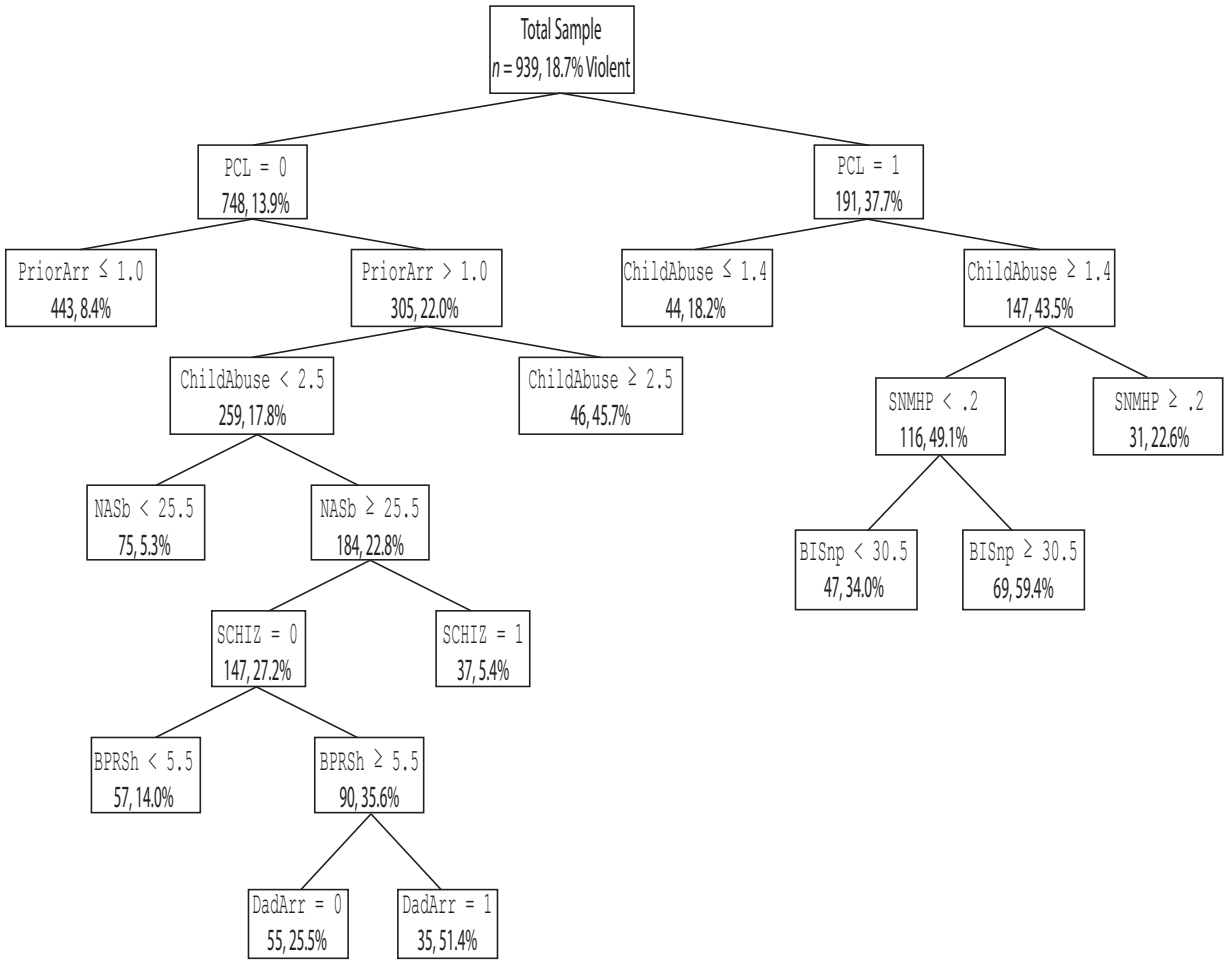

Figure 5.5: Classification tree with an optimal minimum leaf size of 26 and setting cost of false negatives to 1.67 times the cost of false positives.

If we adhered to Berk's (2012) 20:1 false negative to false positive ratio, the resubstitution error (with a minimum leaf size of 30) is .612 but the cross-validated error is .995. Because of these results and the fact that it is impossible to justify a 20:1 ratio for the VRAS definition of violence, this cost ratio is not considered in the remaining analyses.

Suppose we decided not to empirically determine a minimum leaf size but let a min-

imum leaf size of 1, the default setting in MatLab. Doing so, a resubstitution error of .071 is obtained; far better than any other model developed thus far because only 67 of the 939 patients are misclassified. This model is better than the "best" logistic regression or discriminant model examined thus far. The sensitivity of the test is .78, the specificity is .97, the PPV is .86, the NPV is .95, and the AUC is .97. We have yet to come across an actuarial measure this promising. Without cross-validating this model, however, one can blindly believe that an extremely capable model for predicting violence is found; the LOOCV error of .244 provides evidence of this as it is one of the *highest* of any model examined thus far. Based on a cutscore of .37 rather than .50 the model misclassifies only 63 individuals (resubstitution error of .067) but the cross-validation error is .352. This exemplifies the over-fitting of a model and the importance of cross-validation, and provides yet another example of a more flexible model performing well on the data for which the model was fit but far worse on new data.

5.5.4 Ensemble Learning Methods for Decision Trees

A prediction method called *ensemble learning* is now implemented on the VRAS dataset. *Bagging* is an ensemble learning method designed to avoid the overfitting of a model; it is commonly used with classification trees (i.e., tree bagging) (*bag* is a short-hand phrase for bootstrap aggregation; see Breiman, 1996). Suppose a dataset, $\mathbf{X}$, contains n observations. Similar to the bootstrap method, B training sets of size K are generated, where $1 \leq K \leq n$, by randomly sampling (with replacement) from $\mathbf{X}$; each training set is fit by the model and, after aggregating, an average over the B replications provides a predicted response for each observation.

Note that some of the observations in the ith training set, $\mathbf{B}^{(i)}$, may be duplicate observations. The larger K is, the more likely there will be at least one duplicate observation; the probability of such an event is $1 - \frac{n!}{n^K(n-K)!}$. The probability that any given observation is not selected is $(1 - \frac{1}{n})^K$. If $K = n$ and as $n \to \infty$, the probability approaches $e^{-1} \approx .37$.

For a large enough n and when the training sample is equal to n, it would be expected that on average, about 63% of the bootstrap sample consists of unique observations. The 63% represents a probabilistic lower bound; for $K < n$, one would expect more than 63% of the sample to be unique (e.g., in the trivial case where $K = 1$, there are no duplicate observations). The approximately 37% of the observations not used in fitting the model on the ith replication are called *out-of-bag* (OOB); thus, the OOB observations are the testing set and can be used to assess predictive accuracy. For any given observation, by aggregating over the subset of B replications—where the observation was not used to fit the model—the average OOB prediction accuracy can be calculated and this can be compared to the misclassification error; this comparison gives us the average OOB error difference. The OOB errors can also be used to assess the importance of predictors by randomly permuting the OOB data across variables one at a time, and estimating the OOB error after permutation—a large increase in the OOB error indicates the variable's importance in the model.

Random forests (Breiman, 2001) are a tree bagging method that randomly selects a subset of variables at each split. The advantage of randomly selecting variables at each node in the decision tree is that it can decorrelates the trees. This is advantageous because it prevents a single variable from dominating the analysis; for instance, if one predictor is very strong it likely will be the root node for a majority of the trees constructed and the subsequent nodes will be similar as well (i.e., the trees will be highly correlated). Typically, $\sqrt{p}$ predictors are randomly selected for classification trees. Random forests were constructed in MatLab (MATLAB, 2013) and the results are presented in the following sections.

VRAS Random Forest Model in MatLab.

Randomly selecting a subset of the VRAS dataset as the training sample in the random forest algorithm, the remaining observations represent the testing sample. The testing sample contains 30% of the original data (281 observations); the training data contain the remaining 658 observations (the base rate for violence in the training sample is .188, and

.185 in the testing sample). The random forest model was fit to the training sample for $B = 1000$ and a minimum leaf size of 10. After fitting 1000 trees, the random forest model was used to predict violence in the training set (i.e., the observations for initially fitting the model). No nonviolent individual is misclassified as violent but 81.5% of violent individuals are misclassified; thus the model outperforms base rate prediction. The 1000 trees generated can be aggregated to estimate the probability an individual will be violent by computing the proportion of times the individual is classified as violent (an individual is classified as violent if the predicted probability exceeds .50; i.e., costs are considered equal here). This is quite different than probabilities assigned to violence prediction with a single tree; in such instances, the probability an individual will be violent is equal to the proportion of individuals in the terminal node who are violent. These results are summarized in Table 5.12.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	23	0	23
	No ($\bar{B}$)	101	534	635
	Column Totals	124	534	658

Table 5.12: Predicting violence with a random forest model on the training sample. If a patient had a predicted probability greater than .50, a prediction of violence was made; otherwise a prediction of no violence was made.

The parallel coordinates plot (Figure 5.6) displays the predicted probabilities for each of the 658 individuals. The green-colored lines are the individuals who were violent; the blue lines are those who were not. The horizontal axis is the estimated probability an individual will be violent (i.e., the proportion of trees for which an individual was classified as violent). There is not a lot of mixture of green and blue lines, which is ideal. The separation between the two types of individuals implies that the method is performing well.

The results seen thus far are, as noted, based on the training data. The greater concern is with how well violence can be predicted in new observations with the random forest model, as demonstrated with the testing data. Of the 281 observations, only one

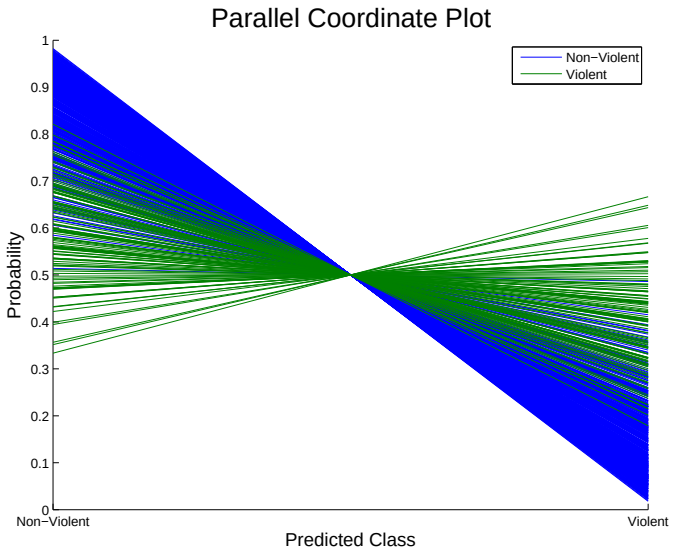

Figure 5.6: Parallel coordinate plot for random forest predictions of violence on the training sample ($n = 658$).

is predicted—incorrectly—to be violent (see Table 5.13). The misclassification error is $^{(1+52)}/_{281} = .189$, slightly more than the base rate. The sensitivity and specificity are, respectively, $^{0}/_{52} = 0$ and $^{228}/_{229} = .996$; the positive and negative predictive values are, respectively, $^{0}/_{1} = 0$ and $^{228}/_{280} = .814$. The parallel coordinate plot in Figure 5.7 indicates that the method is doing only moderately well with a poorer separation of the green and blue lines.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	0	1	1
	No ($\bar{B}$)	52	228	280
	Column Totals	52	229	281

Table 5.13: Predicting violence with a random forest model on the testing sample. If a patient had a predicted probability greater than .50, a prediction of violence was made; otherwise a prediction of no violence was made.

The next analysis is the same as before except that the cost ratio of false negatives to false positives is set to 1.67. At the individual tree level an observation is classified as violent when it belongs to a terminal node where the proportion of violent individuals is

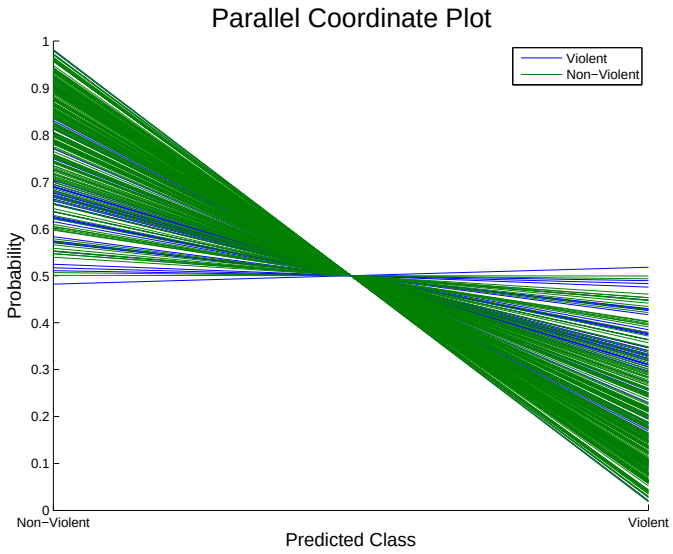

Figure 5.7: Parallel coordinate plot for random forest predictions of violence on the testing sample ($n = 281$).

greater than .37. At the aggregate level (i.e., across all 1000 trees) an individual is predicted to be violent when classified as violent in more than 37% of the trees. The results for the training data (base rate of .187) are displayed in Table 5.14 and Figure 5.8; for the testing data (base rate .189), the results can be found in Table 5.15 and Figure 5.9.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	120	74	194
	No ($\bar{B}$)	3	461	466
	Column Totals	123	535	658

Table 5.14: Predicting violence with a random forest model on the training sample. If a patient had a predicted probability greater than twice the base rate (.37), a prediction of violence was made; otherwise a prediction of no violence was made.

As expected, more predictions of violence are being made. The selection ratio ($P(B)$) for the training data is .295 compared to .035 when the costs were considered equal; for the testing data, the model classifies 24.2% of the sample as violent (compared to 0.4% when costs are equal) and less than half of these are correct. Once again, the model performs well on the training data (resubstitution error .117) but not on the testing data (misclassification

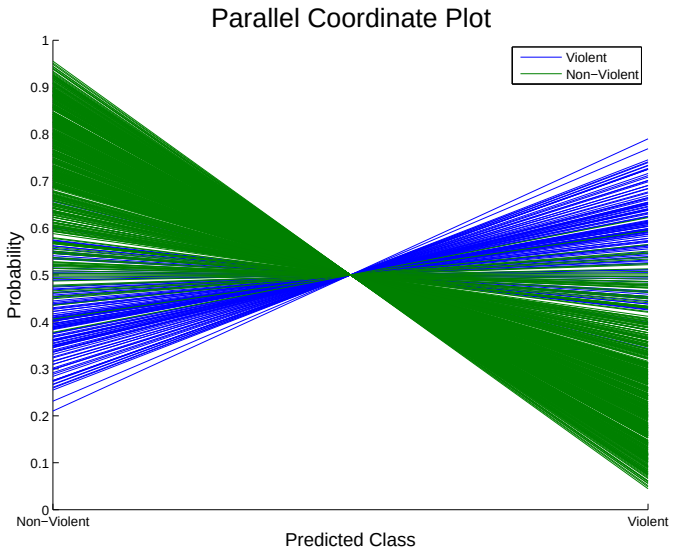

Figure 5.8: Parallel coordinate plot for random forest predictions of violence on the training sample ($n = 658$) with classification cutscore of .37.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	31	37	68
	No ($\bar{B}$)	22	191	213
	Column Totals	53	228	281

Table 5.15: Predicting violence with a random forest model on the testing sample. If a patient had a predicted probability greater than twice the base rate (.37), a prediction of violence was made; otherwise a prediction of no violence was made.

error .210).

Out-of-bag Prediction.

Rather than splitting the data prior to fitting the ensemble method, the entire dataset can be used with cross-validation error estimated from the OOB observations. Relying on OOB observations, the sample size is maximized (i.e., nothing is "wasted") but cross-validated error estimates are still obtained. Carrying this out for the VRAS dataset, the 2×2 contingency Table 5.16 is constructed; the parallel coordinate plot for these results are in Figure 5.10. The parallel coordinate plot shows little separation between the violent and

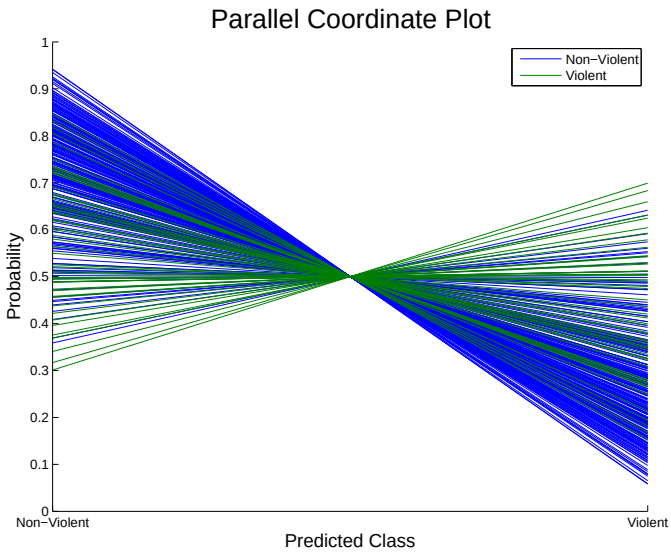

Figure 5.9: Parallel coordinate plot for random forest predictions of violence on the testing sample ($n = 658$) with classification cutscore of .37.

nonviolent individuals. The mean out-of-bag error is .188, slightly above the base rate; the minimum OOB error is .185, slightly less than the base rate.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	4	5	9
	No ($\bar{B}$)	172	758	930
	Column Totals	176	763	939

Table 5.16: Predicting violence with a random forest model on the entire sample. If a patient had a predicted probability greater than .50, a prediction of violence was made; otherwise a prediction of no violence was made.

Running the same analysis but setting the classification cutscore to .37 produces similar results. As seen in Table 5.17, the model classifies 32.1% of individuals as violent, only 37.9% of whom are. The parallel plot (Figure 5.11) illustrates the model's lack of differentiation among violent individuals.

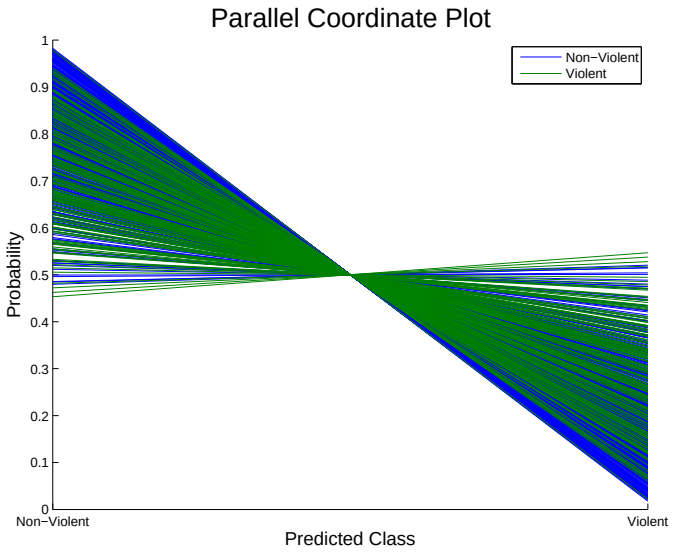

Figure 5.10: Parallel coordinate plot for out-of-bag random forest predictions on the entire sample ($n = 939$).

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	114	187	301
	No ($\bar{B}$)	62	576	638
	Column Totals	176	763	939

Table 5.17: Predicting violence with a random forest model on the entire sample. If a patient had a predicted probability greater than twice the base rate (.37), a prediction of violence was made; otherwise a prediction of no violence was made.

Variable Selection.

Out-of-bag observations allow the quantification of variable importance to the classification trees. Thus far, the decision trees studied have included all thirty-one variables. For each variable, the values are randomly permuted and the increase (or decrease) in the OOB error calculated (i.e., the difference in OOB error before and after permutation). This is carried out for every tree and normalized with the standard deviations of the differences. Variables with larger average differences can be quantified as more important than variables with smaller averages. A bar plot of the variable importance is in Figure 5.12.

From Figure 5.12 several variables appear to be more important than others and

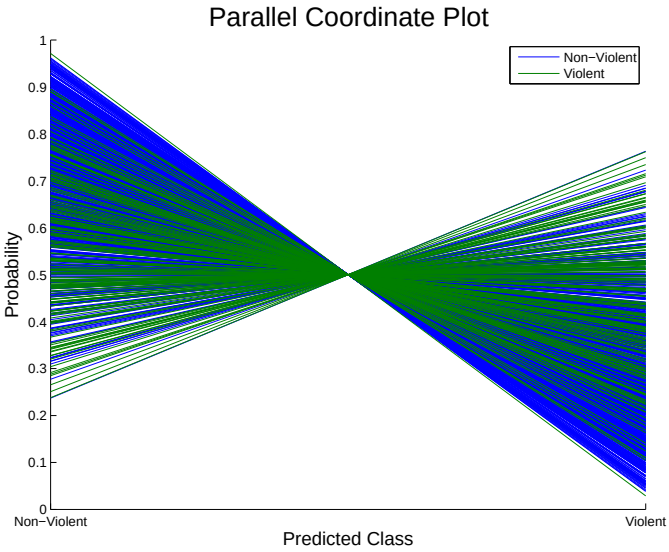

Figure 5.11: Parallel coordinate plot for out-of-bag random forest predictions on the entire sample ($n = 939$) with a cutscore of .37.

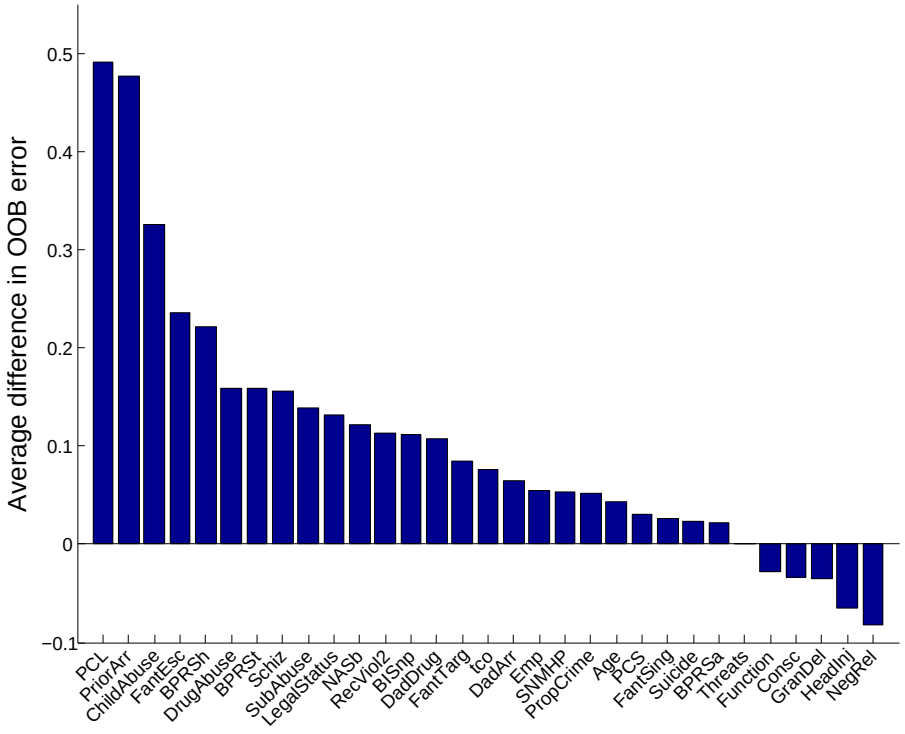

Figure 5.12: Measure of importance for each of the thirty-one variables. The importance measure is the average of the differences in out-of-bag error before and after permuation across all trees.

several actually decrease the OOB error after permutation. The variables that appear the most important are `PCL` and `PriorArr`. It is interesting to note that `Schiz` is among the more important variables and `Age` is not. The decision for removing variables will be conservative; only variables with average OOB error differences near and less than zero are removed (`Threats`, `Function`, `Consc`, `GrandDel`, `HeadInj`, `NegRel`); thus, the final model consists of 25 variables. When applying the cost function, all variables have negative mean OOB error differences implying that the OOB error *decreases* on average after random permutation for all the variables in this model.

Final Model.

For the final model, the data are split into a training set and a testing set. Again, the training set contains 658 of the original 939 observations and the testing set the remaining 281. The base rate for violence in the training set is .1884; in the testing set, .1851. The final model was estimated with 10000 trees omitting the variables discussed in the previous section. The mean, median, minimum, and maximum OOB errors are, respectively, .1873, .1869, .1763, and .2097, slightly better than with all the variables. Based on the model with the testing data and a cutscore of .50, only two individuals are predicted to be violent, only one of which is (see Table 5.18). The parallel coordinate plot (Figure 5.13) appears to indicate that the final model does a slightly better job than before.

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	1	1	2
	No ($\bar{B}$)	51	228	279
	Column Totals	52	229	281

Table 5.18: Predicting violence with the final random forest model on the testing sample. If a patient had a predicted probability greater than .50, a prediction of violence was made; otherwise a prediction of no violence was made.

The ROC plot is given in Figure 5.14; the AUC for the final model is .748. The cost

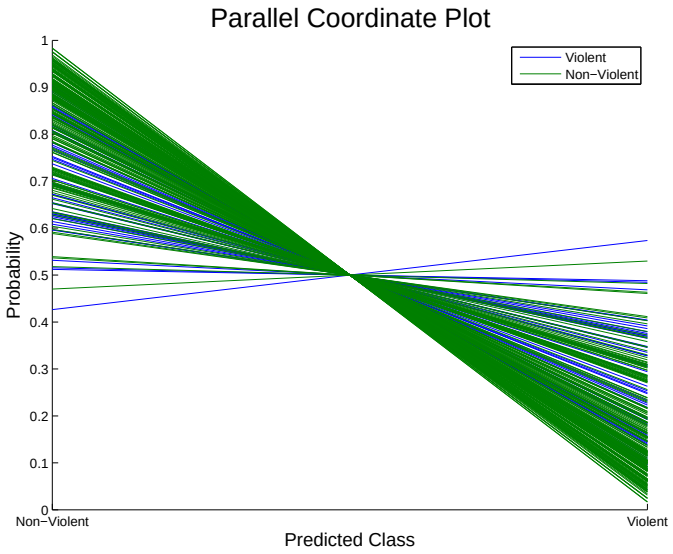

Figure 5.13: Parallel coordinate plot for final random forest model predictions with the testing data.

function was not applied in fitting the final model, but a .37 cutscore can be applied at the aggregate-tree level to classify individuals as violent or not. So doing, nineteen individuals are classified as violent, ten of whom are.

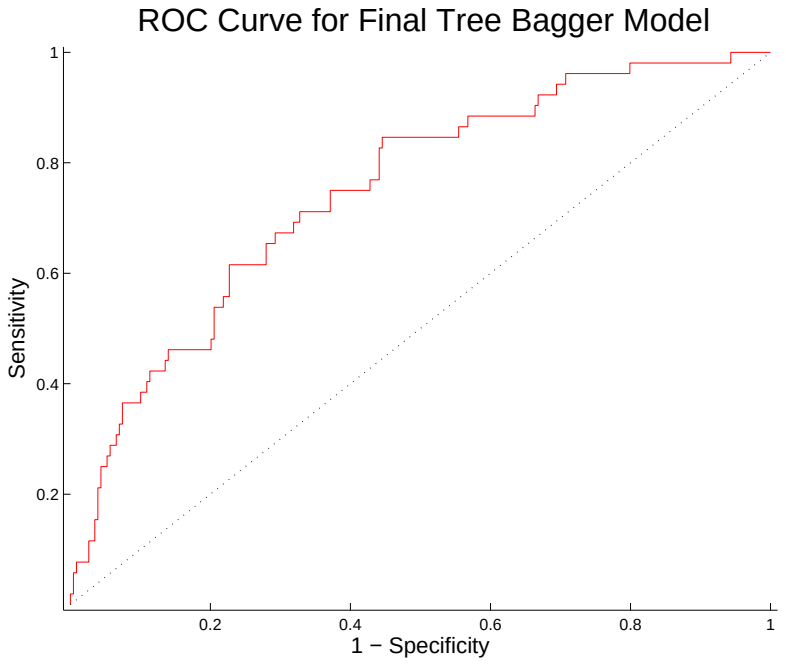

Figure 5.14: ROC plot for the final random forest model. The AUC is .75.

The .37 cutscore was arbitrarily chosen; however, an optimal cutscore can be determined through a performance curve similar to the ROC curve that plots the expected cost against possible cutscores. Figure 5.15 displays the estimated cost function; it suggests that the cost is minimized when the cutscore is approximately .37 (slightly below the base rate when carried to more decimal places). Based on this cutscore, the model predicts twenty-three individuals to be violent (about 8% of the sample), ten of them incorrectly. Although the minimum cost is at the .37 cutscore, cutscores slightly below and much higher than .37 do not produce much greater overall costs; in other words, the cost function appears to approach an asymptote suggesting that around a .32 cutscore a trade-off between false negatives and false positives commences. This is indeed the case as Table 5.19 demonstrates. The first and third columns give the actual outcome of the individual (1 = Violent, 0 = Nonviolent); the second column is the predicted probability for violence (i.e., the proportion of trees for which the individual was classified as violent). Only individuals with estimated probabilities greater than .32 are provided. The pattern roughly switches between violent and nonviolent.

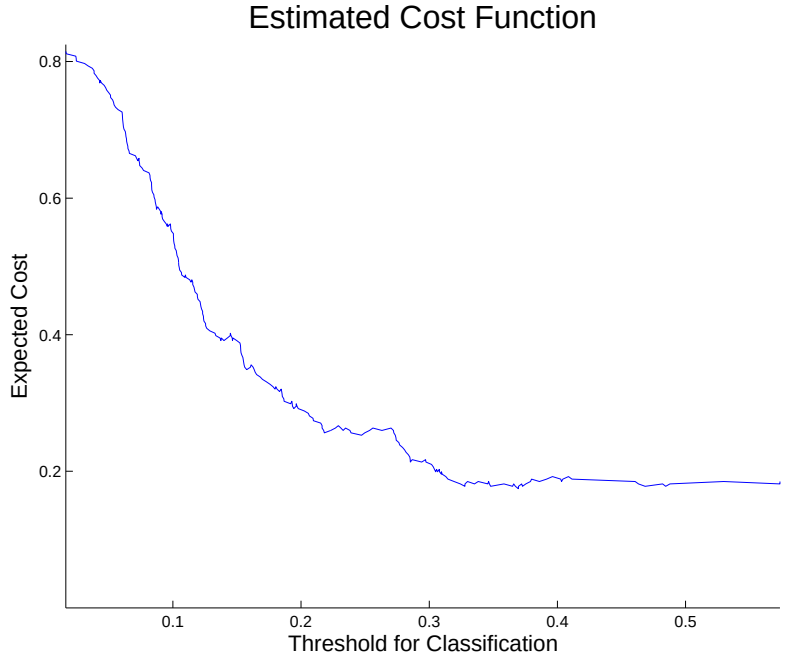

Figure 5.15: Estimated cost function for the final random forest model. The estimated model cost (assuming equal costs of false positives and negatives) is plotted across all cutscores.

Actual Outcome	Predicted Probability	Actual Outcome	Predicted Probability
0	.324	1	.379
1	.328	0	.380
1	.328	1	.386
0	.330	1	.392
1	.335	0	.396
0	.338	0	.403
1	.346	1	.403
0	.346	1	.404
0	.347	0	.409
1	.348	0	.411
0	.358	0	.461
1	.365	0	.463
0	.366	1	.468
0	.367	0	.482
1	.370	1	.485
1	.370	1	.488
0	.372	0	.530
1	.373	1	.574
1	.375		

Table 5.19: Predicted probabilities and actual outcomes for individuals with a predicted probability greater than .32 based on the final random forest model.

5.6 Conclusion

The results given in this chapter reiterate the argument that predicting violent behavior is extremely difficult. Unless unequal costs regarding false positives and negatives are assumed—particularly when false negatives are considered to be more costly than false positives—our methods typically make a small number of predictions of violence. Thus, unless the model explicitly states false negatives as more costly than false positives, predictions of violence will be infrequent. G. T. Harris and Rice (2013) claim "it can be reasonable for public policy to operate on the basis that a miss (e.g., failing to detain a violent recidivist beforehand) is twice as costly as a false alarm (e.g., detaining a violent offender who would

not commit yet another violent offense)" (p. 106). Although this may be true, it is ethically questionable to assume that costs are anything but equal unless a public policy explicitly states otherwise.

The analyses presented also demonstrate the importance of cross-validation. Without cross-validating a model, results and conclusions regarding accuracy can be misleading and an extreme overestimate. For each type of model constructed (logistic regression, discriminant analysis, classification trees), when assessed on the data used to fit it, performed quite well, almost always outperforming base rate prediction (although as noted, this should be a minimal requirement). But when cross-validation methods were implemented, the results were dramatically different. Rarely did the model outperform base rate prediction on the testing sample, and the resubstitution error was higher. Table 5.20 is from the VRAS study (derived from Table 6.7 in Monahan et al., 2001). When individuals who fall into the very high- and high-risk groups are classified as violent and all others as nonviolent, the model correctly classifies 86.0% of individuals which is better than nearly every model found in these analyses. Given this cutscore, the model has a sensitivity and specificity of, respectively, $^{(48+57)}/_{176} = .597$ and $^{(135+229+339)}/_{763} = .921$; the positive and negative predictive values are, respectively, $^{(48+57)}/_{(63+102)} = .636$ and $^{(135+229+339)}/_{(183+248+343)} = .908$. Recall from Chapter 1 that the COVR was a combination of ten ICT models, five of which were kept. The authors claim that this "multiple model approach minimizes the problem of data overfitting that can result when a single 'best' model is used" (p. 127). Because the authors did not do a cross-validation, it is not possible to determine how much the model overfits the data, but it certainly seems that it does. As McCusker (2007) says,

> One could wonder whether the iterative classification tree methodology (a technique that involves repetitive sifting of risk factors) that was used to create the COVR ended up, in a sense, fitting the data in the development sample too specifically. Perhaps as a very carefully tailored garment will be expected to fit one individual perfectly but most other people not as well, so the algorithms of the COVR ought

to be anticipated to classify other samples less exactly than they categorized the members of the development sample. (p. 682)

		Violence		
		Yes	No	Row Totals
Risk Group	Very High	48	15	63
	High	57	45	102
	Average	48	135	183
	Low	19	229	248
	Very Low	4	339	343
	Column Totals	176	763	939

Table 5.20: COVR risk groups from Monahan et al. (2001, cf. Table 6.7, p. 125).

As important as cross-validation is when developing a model, the true test lies in how well the model does with an independent sample. Therefore, the next step in validating the model is to assess the model's accuracy with a completely new sample. Chapter 7 looks at the performance of the COVR across six different studies (we also examine the predictive ability of the Static-99 and Static-2002).

A final point is to note that in general, as the model became more flexible (e.g., moving from a linear to a quadratic classifier, or decreasing the minimum leaf size) the resubstitution error decreased but the cross-validated error increased. This is an example of the bias-variance trade-off discussed in Chapter 6.

Chapter 6

The Variance-Bias Trade-off

"Strategy is about making choices, trade-offs"

— Michael Porter

The mean squared error of a predictive model can be decomposed into three parts: the variance of the model, the squared bias, and irreducible error. The trade-off between variance and bias in predictive modeling is an important concept commonly overlooked; often, test error is not minimized when using unbiased predictors. Trading off bias for variance (i.e., increasing squared bias while decreasing variance) will lead to decreased test error, a desired result in prediction. This chapter looks at the variance-bias trade-off and examines several shrinkage estimators designed to reduce overall error at the cost of increased bias; the biased estimators that are discussed consistently outperform their unbiased counterparts in terms of mean squared error.

6.1 Introduction

Let $x = (x_1, x_2, \ldots, x_p)$ be a collection of p predictor variables and let Y be a response variable. Suppose

$$Y = f(x) + \varepsilon,$$

where ε is an error random variable with $\mathbb{E}(\varepsilon|X) = 0$ and $\mathbb{V}(\varepsilon|X) = \sigma_\varepsilon^2$; that is, there exists a function $f(\cdot)$ for modeling the response variable, plus unknown error ε.

A *loss function* measures the error between the estimated function, $\hat{f}(x)$, and the true one, $f(x)$; a common type of loss function is the *squared error loss*,

$$\mathcal{L}_2(f(x), \hat{f}(x)) = \left(\hat{f}(x) - f(x)\right)^2.$$

The quality of an estimator can be quantified by taking the expectation of the loss function; this is called the *risk function* and defined as

$$\mathcal{R}(f(x), \hat{f}(x)|X) = \mathbb{E}(\mathcal{L}(f(x), \hat{f}(x))|X).$$

The interest in prediction reduces to estimating the function, $f(x)$, given the observed data, $(\mathbf{X}, \mathbf{y})$, where $\mathbf{X}$ is an $n \times p$ matrix of n observations on p predictors and $\mathbf{y}$ is an $n \times 1$ vector containing the outcome; the estimated function will be denoted as $\hat{\mathbf{y}} \equiv \hat{f}(x) \equiv \hat{f}(x|\mathbf{X})$. Given the observed data, $(\mathbf{X}, \mathbf{y})$, the estimated risk function for the estimator $\hat{\mathbf{y}}$ using the squared loss function, as

$$\frac{1}{n}\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where $\|\cdot\|_2$ is the Euclidean, or ℓ_2, norm (for simplicity, the subscript 2 will be dropped); this is commonly known as the *mean squared error* and denoted MSE.

6.1.1 Variance-Bias Decomposition

The risk function for the squared error,

$$\mathcal{R}(y, \hat{y}|X) = \mathbb{E}\left(\mathcal{L}_2(y, \hat{y})\,|X\right) = \mathbb{E}\left[(y - \hat{y})^2|X\right],$$

is decomposable into two parts (see Chapter 5; for proof, see Appendix A):

$$\mathcal{R}(y, \hat{y}|X) = \mathbb{E}\left[(f(x) - \hat{y})^2\big|\, X\right] + \mathbb{V}(\boldsymbol{\varepsilon}|X). \tag{6.1}$$

The first term, $\mathbb{E}\left[(f(x) - \hat{y})^2 | X\right]$, is the *reducible error*. The better the estimator $\hat{y}$, the closer the reducible error is to zero; the reducible error is equal to zero when $\hat{y} = f(x)$. The second term in Equation (6.1), $\mathbb{V}(\boldsymbol{\varepsilon}|X)$, is the *irreducible error* and represents a lower bound for the risk function. Even when $f(x)$ is perfectly estimated (i.e., $\hat{y} = f(x)$), $\mathcal{R}(y, \hat{y}) = \mathbb{V}(\boldsymbol{\varepsilon}|X) > 0$.

The reducible error can be further decomposed (see Appendix A) so that the risk function reduces to

$$\mathcal{R}(y, \hat{y}|X) = \left(\text{bias}(\hat{y}|X)\right)^2 + \mathbb{V}(\hat{y}|X) + \mathbb{V}(\boldsymbol{\varepsilon}|X). \tag{6.2}$$

The first term is the squared bias associated with the estimator for the function. The bias measures how much, on average, the estimated function over- or underestimates the true function, $f(x)$; an unbiased estimator has a squared bias equal to zero. The second term is the variance of the estimator and represents how much, on average, the estimated function deviates from the true function. Thus, reducible error depends on both the variance and the bias of the estimator.

6.1.2 Brier Score

Let $\mathbf{y}_i$ be an outcome vector of size K for an ith observation corresponding to one of K classes that the outcome may belong to. Explicitly, this outcome of being in the kth class can be defined as $\mathbf{y}_i = \mathbf{e}_k$, where $\mathbf{e}_k$ is a $K \times 1$ vector with the kth entry equal to one and all others equal to zero:

$$y_{ik} = \begin{cases} 1 & \text{if the } i\text{th observation belongs to class } k, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\hat{\mathbf{y}}_i$ be an estimator for $\mathbf{y}_i$ where $\hat{\mathbf{y}}_i$ is a $K \times 1$ vector containing estimated probabilities that the ith observation belongs to each of the K classes: $\hat{\mathbf{y}}_i = (\hat{y}_{i1}, \ldots, \hat{y}_{iK})^T$, where

$\sum_{k=1}^{K} \hat{y}_{ik} = 1$. For instance, consider the outcome of violent behavior so that $\mathbf{y}_i = (1,0)^T$ when violent behavior is observed for the ith individual and $(0,1)^T$ otherwise; $\hat{\mathbf{y}}_i$ contains the estimated probabilities assigned to the ith individual (i.e., $\hat{y}_{i1}$ is the estimated probability of violent behavior and $\hat{y}_{i2}$, for non-violent behavior).

Given n observations, Brier (1950) devised a *verification score* defined as

$$\text{VS} = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|^2 = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T(\mathbf{y}_i - \hat{\mathbf{y}}_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} (y_{ik} - \hat{y}_{ik})^2. \tag{6.3}$$

This measure is commonly referred to as the *Brier score* in the literature and is a variant of the mean squared error.

Clearly the minimum of VS is 0, occurring when the estimated probability corresponding to the true class is equal to 1 for all observations. The maximum of VS is 2, occurring when the estimated probability for one of the incorrect classes is equal to one for all observations.

Brier (1950) notes that when given prior probabilities for each class, say $p_1, p_2, \ldots, p_K$ where $p_1 + \cdots + p_K = 1$, naïve prediction is to let $\hat{y}_{ik} = p_k$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ (i.e., prediction using the base rates). The prior probability can be estimated as

$$\hat{p}_k = \frac{1}{n}\sum_{i=1}^{n} y_{ik}.$$

In such a scenario, the Brier score is

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} (y_{ik} - \hat{y}_{ik})^2 &= \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} (y_{ik}^2 - 2y_{ik}p_k + p_k^2) \\
&= \frac{1}{n}\sum_{k=1}^{K}\left(\sum_{i=1}^{n} y_{ik}^2 - 2p_k\sum_{i=1}^{n} y_{ik} + \sum_{i=1}^{n} p_k^2\right) \\
&= \frac{1}{n}\sum_{k=1}^{K}\left(np_k - 2np_k^2 + np_k^2\right)
\end{aligned}$$

$$= \sum_{k=1}^{K} \left(p_k - p_k^2\right)$$
$$= 1 - \sum_{k=1}^{K} p_k^2.$$

Comparing a Brier score obtained using a predictive tool to the naïve method above is one way to quantify the predictive tool's incremental value. The maximum of the above is when $p_1 = \cdots = p_K = K^{-1}$ in which case the minimal Brier score is $1 - 1/K$; the minimum is 0 when $p_k = 1$ for some k and $p_{k'} = 0$ for all $k' \neq k$. Note the similarities between the Brier score using naïve prediction and the Gini index discussed in Chapter 1 for determining the purity of a terminal node in a decision tree.

For $K = 2$ classes, $\hat{y}_{i2} = 1 - \hat{y}_{i1}$; thus, the Brier score is equal to

$$\frac{1}{n} \sum_{i=1}^{n} 2(y_{i1} - \hat{y}_{i1})^2.$$

Because $p_2 = 1 - p_1$ the minimal Brier score obtained with naïve prediction using the base rates is $2p_1(1 - p_1)$; thus, a base rate closer to one half implies a larger minimal Brier score.

Alternatively, for $K = 2$ classes the Brier score can be written as

$$\text{VS}' = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{6.4}$$

where y_i is the outcome of interest for the ith observation (e.g., $y_i = 1$ if the ith individual is violent and 0 otherwise) and $\hat{y}_i$ is the estimated probability of the outcome for the ith observation (e.g., the probability the ith individual is violent). The Brier score still has a minimum of 0 corresponding to perfect prediction, but the maximum is now one.

Brier Score Decomposition.

Because violent and dangerous behavior is primarily considered a binary outcome, the focus will be restricted to such cases. Consider the Brier score given by Equation (6.3)

and suppose there exist R distinct probability estimates. Let $\hat{\mathbf{y}}^{(r)}$, $r = 1, \ldots, R$, represent a collection of R mutually exclusive subsets for the estimated probabilities; thus, for $r = 1, \ldots, R$, $\hat{\mathbf{y}}^{(r)}$ represents a unique estimated probability vector of size 2×1. Within each subset, n_r probabilities are estimated. For example, consider the risk categories given by the Classification of Violence Risk (COVR) shown in Table 6.1. There are $R = 5$ estimated probabilities associated with the five risk categories; thus, $\hat{\mathbf{y}}^{(5)} = (.76, .24)^T$, $\hat{\mathbf{y}}^{(4)} = (.56, .44)^T$, $\hat{\mathbf{y}}^{(3)} = (.26, .74)^T$, $\hat{\mathbf{y}}^{(2)} = (.08, .92)^T$, $\hat{\mathbf{y}}^{(1)} = (.01, .99)^T$.

Category	Risk	Probability
5	Very High	.76
4	High	.56
3	Average	.26
2	Low	.08
1	Very Low	.01

Table 6.1: The five risk categories for the Classification of Violence Risk (COVR) with estimated probabilities of violence (Monahan et al., 2001).

For a given subset, the Brier score is

$$\mathrm{VS}^{(r)} = \frac{1}{n_r} \sum_{j=1}^{n_r} \|\mathbf{y}_j^{(r)} - \hat{\mathbf{y}}^{(r)}\|^2,$$

where $\mathbf{y}_j^{(r)}$ is the 2×1 outcome for the jth individual in the rth subset. Murphy (1972a) shows (see also Murphy, 1972b; Sanders, 1958, 1963) that the Brier score for a given subset can be decomposed as

$$\frac{1}{n} \sum_{r=1}^{R} n_r \|\hat{\mathbf{y}}^{(r)} - \bar{\mathbf{y}}^{(r)}\|^2 + \frac{1}{n} \sum_{r=1}^{R} n_r \langle \bar{\mathbf{y}}^{(r)}, \mathbf{e} - \bar{\mathbf{y}}^{(r)} \rangle, \tag{6.5}$$

where $\mathbf{e}$ is an 2×1 vector of 1's, $\bar{\mathbf{y}}^{(r)} = \frac{1}{n_r} \sum_{j=1}^{n_r} \mathbf{y}_j^{(r)}$, and $\langle \cdot, \cdot \rangle$ is the inner product, or dot product, in the Euclidean space (see Appendix A for a proof of this decomposition).

The first term in Equation (6.5) is commonly referred to as the *reliability* or *calibration*

of a predictive measure (note the counterintuitive result that a lower reliability is better); the second term has been called the *resolution* or *refinement* and provides an idea of the uncertainty in prediction. Ferri, Hernández-Orallo, and Flach (2011) demonstrate the use of the Brier score decomposition in creating *Brier curves*; such a curve is meant to be an alternative to the typical ROC curve, with the area under the Brier curve equal to the Brier score.

The Brier score decomposition is analogous to the sum of squares error decomposition where the sum of squares error is decomposed into the *lack of fit* and *pure error*. Both decompose a measure of error into reducible error (reliability, lack of fit) and irreducible error (resolution, pure error).

Murphy (1973) further decomposed the Brier score into three partitions; the first partition (the reliability) remains the same; the second partition, however, is further decomposed into two parts so that the Brier score now reduces to

$$\frac{1}{n}\sum_{r=1}^{R} n_r \|\hat{\mathbf{y}}^{(r)} - \bar{\mathbf{y}}^{(r)}\|^2 - \frac{1}{n}\sum_{r=1}^{R} n_r \|\bar{\mathbf{y}}^{(r)} - \bar{\mathbf{y}}\|^2 + \langle \bar{\mathbf{y}}, \mathbf{e} - \bar{\mathbf{y}} \rangle, \tag{6.6}$$

where $\bar{\mathbf{y}} = \frac{1}{n}\sum_{r=1}^{R}\sum_{j=1}^{n_r} \mathbf{y}_j^{(r)}$ (see Appendix A for a proof). Murphy (1973) refers to the three terms respectively as *reliability*, *resolution*, and *uncertainty*. Reliability, as mentioned, is the same as in Equation (6.5) and represents the reducible error. The uncertainty term is

$$\langle \bar{\mathbf{y}}, \mathbf{e} - \bar{\mathbf{y}} \rangle = \bar{\mathbf{y}}^T \mathbf{e} - \bar{\mathbf{y}}^T \bar{\mathbf{y}} = 1 - (p_1)^2 + (1 - p_1)^2 = 2p_1 (1 - p_1),$$

the minimal Brier score obtained using naïve base rate prediction; geometrically it is proportional to the cosine of the angle, θ, of the mean vector, $\bar{\mathbf{y}}$, and the vector $\mathbf{e} - \bar{\mathbf{y}}$. When $p_1 = p_2$, $\theta = 0$ and $\cos(\theta) = 1$; when $p_1 = 1$ or $p_2 = 1$, $\theta\pi/2$ and $\cos(\theta) = 0$. Thus, the uncertainty depends solely on the base rate and is equal to the minimal Brier score using naïve prediction. The resolution measures the difference in the conditional prior probabilities from

the overall prior. The larger the resolution the better; the minimum is 0 when the base rate probability is used for the estimated probabilities. The maximum of the resolution is equal to the uncertainty (consequently, the Brier score will equal the reliability); this occurs when the conditional probabilities are all one or zero. Stated alternatively, the resolution can be thought of as how much the prediction tool improves accuracy over naïve prediction.

Example Using the VRAG.

Table 6.2 displays the nine risk categories for the Violence Risk Appraisal Guide (VRAG) and their estimated probabilities for recidivism; there are $R = 9$ unique probability estimates. Kröner, Stadtland, Eidt, and Nedopil (2007) examined the predictive validity of the VRAG in a German sample; the sample consisted of 113 men and women previously charged with a criminal offense and who were evaluated at a forensic psychiatric unit during the years 1994–1995. The outcome variable is recidivism, determined using official records. The authors asserted that the VRAG was successfully replicated in the German sample based on the area under the ROC curve (AUC = .70). Table 6.2 displays the estimated frequencies based on the authors results (derived from the proportions given in Table 4, p. 94).

Category	VRAG Score	Probability	Violent	Nonviolent	Total
9	≥ 28	1.00	2	0	2
8	$[21, 27]$	.76	0	1	1
7	$[14, 20]$	.55	8	2	10
6	$[7, 13]$	.44	8	5	13
5	$[0, 6]$	.35	8	16	24
4	$[-7, -1]$	.17	8	19	27
3	$[-14, -8]$	.12	7	19	26
2	$[-21, -15]$	.08	1	8	9
1	≤ -22	.00	0	1	1

Table 6.2: Observed violent and nonviolent individuals for each of the nine VRAG categories derived from Kröner et al. (2007, Table 4, p. 94); probabilities are in reference to recidivism (G. T. Harris et al., 1993).

The observed Brier score for these data is $\text{VS}' = .21$ (the Brier score provided here uses Equation (6.4), so the maximum is one). An `R` function for producing the Brier score and its parts for a two-category outcome variable is provided in Appendix C. Decomposing the score into its three parts gives the following:

$$\begin{aligned}\text{VS}' &= \text{Reliability} - \text{Resolution} + \text{Uncertainty} \\ &= .02 - .04 + .23.\end{aligned}$$

It was noted earlier that the maximum of the resolution is equal to the uncertainty (the minimum is zero). The ratio of the resolution to uncertainty provides a way to measure how much the predictive tool improves naïve prediction. The uncertainty and resolution will not change when the estimated probabilities change because the two parts make up the irreducible error for the data; thus, the ratio is inherent to the data, given the probabilities assigned by the predictive tool. The VRAG accounts for about $^{\text{Resolution}}/_{\text{Uncertainty}} = .18$ of the uncertainty. Using naïve prediction (i.e., assigning an estimated probability of violence to each individual equal to the base rate, .37), the Brier score is equal to Uncertainty $= .23 = p_1(1 - p_1)$; thus, the VRAG improves prediction over naïve use of base rates but only slightly ($\text{VS}' = .21$ compared to Uncertainty $= .23$).

6.2 The Variance-Bias Trade-off

As noted in Chapter 5, as some models became more flexible (e.g., reducing the minimum leaf size in a decision tree), the training error decreased but the testing error increased. As Equation (6.2) suggests, this increase in testing error can be attributed to either an increase in the variance of the predictor or an increase in the squared bias, or both. Higher variance implies that a change in a set of observations (e.g., applying a model to a new data set) can lead to dramatic changes in the model error; a higher bias implies that the

model assumes a less complex relationship than is true (e.g., assuming a linear relationship when the true relationship is nonlinear). An increase in model flexibility generally leads to an increase in the variance and a decrease in the squared bias (Hastie et al., 2009, p. 38). The test error may initially decrease as the model becomes more flexible but it eventually increases; in contrast, the training error always decreases because the more flexible model fits the data more closely. Fitting too complex of a model leads to increased test error and can be thought of as overfitting; similarly, fitting models leading to decreased test error can be considered underfitting. Ideally, the researcher wants to find the minimal point where the model neither under- nor ovefits the data; this minimum can be estimated using cross-validation.

As an illustration, consider the function

$$f(x) = 5 - 3x + 2x^2 - 7x^3 + \varepsilon;$$

one-hundred observations were simulated from the above function, where $\mathbb{V}(\varepsilon|X) = 1$, and plotted in Figure 6.1; the black curve represents the true cubic function. Figure 6.2 plots the linear, quadratic, cubic, and quartic least-squares line fit to the data.

Test error is estimated using leave-one-out cross-validation (LOOCV) (see Chapter 5) and plotted against different polynomial regression models (ranging from first-order up to twelfth-order; see Figure 6.3). The horizontal dashed line represents $\mathbb{V}(\varepsilon|X) = 1$, the irreducible error. The training error (blue line) approaches this line quickly and eventually decreases to 0. The testing error is at a minimum for the cubic model, as expected. The test error for the fourth-order model is only slightly larger; but moving to a tenth- or higher-order model leads to large increases in the testing error.

The test error can be split into the squared bias and the variance of the model and this is shown in Figure 6.4. The squared bias is at a maximum for the linear model and is still large for the quadratic model but quickly decreases for the cubic model. Further

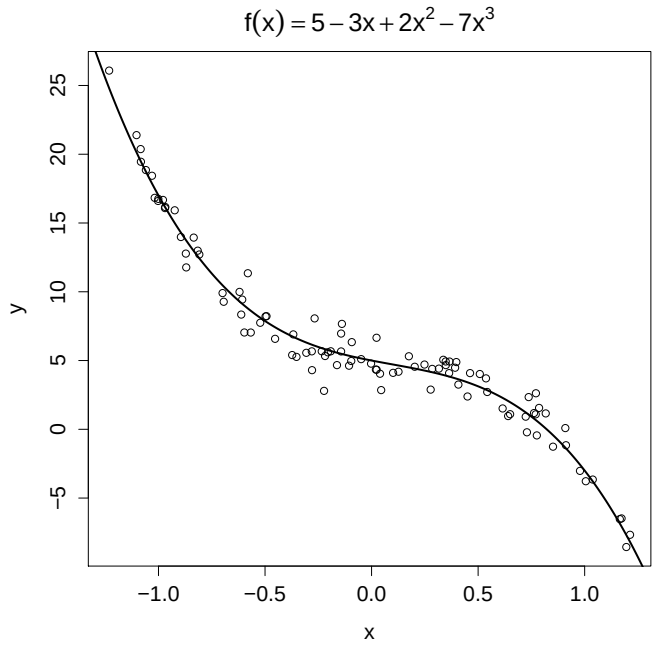

Figure 6.1: Scatter plot of data simulated from $f(x) = 5 - 3x + 2x^2 - 7x^3 + \varepsilon$ where $\mathbb{V}(\varepsilon|X) = 1$. The black line represents the true relationship.

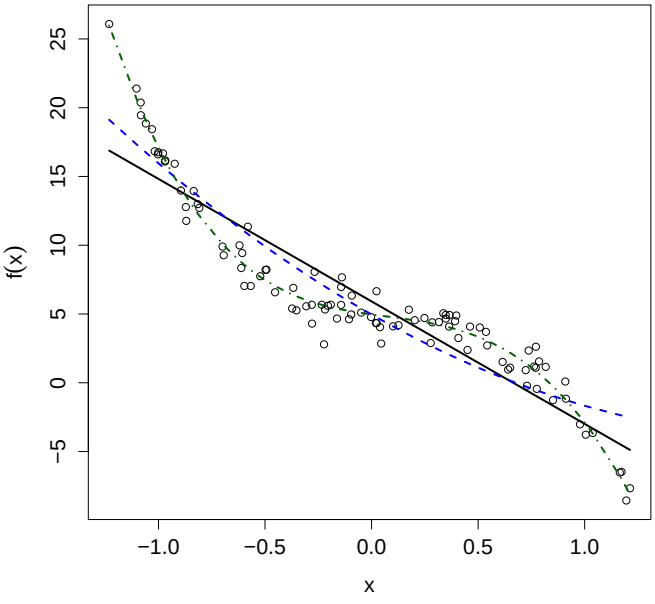

Figure 6.2: Scatter plot of data simulated from $f(x) = 5 - 3x + 2x^2 - 7x^3 + \varepsilon$ where $\mathbb{V}(\varepsilon|X) = 1$. The lines represent the different least squares fits.

increasing the flexibility of the model leads to near-zero bias. In contrast, the variance is near zero for the linear model and remains near zero until the tenth-order model when the variance rapidly increases with greater flexibility.

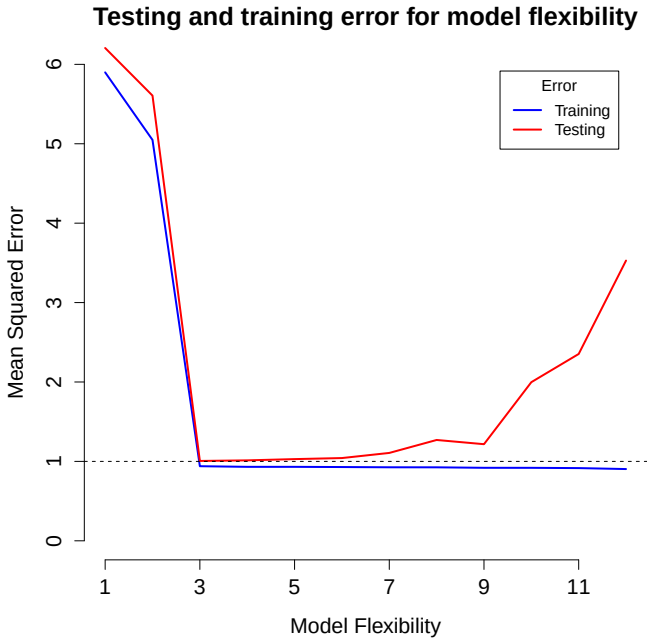

Figure 6.3: Training and testing error plotted against different levels of model flexibility (i.e., different order polynomial regression models).

It is clear that the linear and quadratic models underfit the data and the quartic- and higher-order models overfit the data (although this overfitting is hardly noticeable until the degree is greater than or equal to 10). This simple example illustrates the importance of variance-bias trade-off; in building the best predictive model, often one must sacrifice an increase in bias for less variance to minimize the overall test error.

6.3 Shrinkage Estimators

Shrinkage estimators are designed to improve an estimator by "shrinking" the estimator toward zero or some other value (e.g., the mean of the variable being estimated). Often this shrinkage introduces or increases bias but reduces the variance, and consequently,

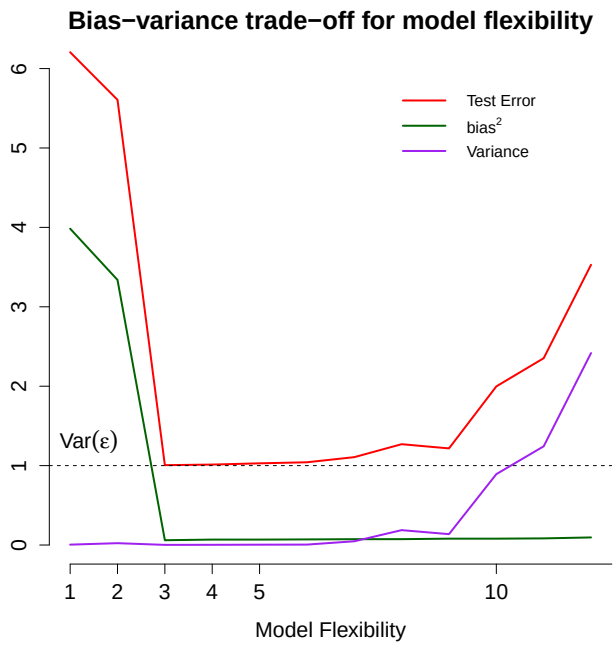

Figure 6.4: Testing error broken down into its two main components, $\Big(\text{bias}(\hat{f}(x))\Big)^2$ and $\mathbb{V}(\hat{f}(x)|X)$, across different levels of model flexibility.

may provide an estimator with smaller overall test error.

Several type of shrinkage estimators are presented in the following sections. The chapter concludes with a demonstration using the MacArthur Violence Risk Assessment Study (Monahan et al., 2001) fit with a main effects logistic regression model.

6.3.1 $n+1$ Estimator

Suppose one is interested in estimating the population variance, σ_X^2. The common choice for an estimator is

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean and n is the number of observations. The estimator is unbiased (i.e., $\mathbb{E}(\hat{\sigma}^2) = \sigma_X^2$) but it is not a unique unbiased estimator; in fact, there exist an infinite number of unbiased estimators of the form $\hat{\sigma}^2 = \sum_{i=1}^{n} c_i(X_i - \bar{X})^2$ where $\sum_{i=1}^{n} c_i = {}^{n}/_{(n-1)}$. The unbiased estimator in the above equation does, however, have minimal

variance among all unbiased estimators of the population variance; thus, it is the uniformly minimum-variance unbiased estimator (UMVUE). It can be shown that the variance and, because the estimator is unbiased, the mean squared error of the UMVUE estimator is

$$\mathrm{MSE}_{\text{UMVUE}} = \mathbb{V}\left(\hat{\sigma}^2_{\text{UMVUE}}\right) = \frac{2\sigma_X^4}{(n-1)},$$

Another commonly used estimator of σ_X^2 is the maximum likelihood (ML) estimator,

$$\hat{\sigma}^2_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

This is a biased estimator and it underestimates the population variance:

$$\text{bias}\left(\hat{\sigma}^2_{\text{ML}}\right) = \mathbb{E}\left(\hat{\sigma}^2_{\text{ML}}\right) - \sigma_X^2 = -\frac{1}{n}\sigma_X^2.$$

However, the variance of the ML estimator,

$$\mathbb{V}\left(\hat{\sigma}^2_{\text{ML}}\right) = \frac{2(n-1)\sigma_X^4}{n^2} < \frac{2\sigma_X^4}{(n-1)},$$

and the mean square error,

$$\mathrm{MSE}_{\text{ML}} = \frac{(2n-1)\sigma_X^4}{n^2} < \frac{2\sigma_X^4}{(n-1)},$$

when $n > 1$.

Hubert (1972) points out that an even better estimator in terms of variance is the estimator,

$$\hat{\sigma}^2_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

which we refer to as the $n+1$ estimator. Again the estimator underestimates the true

variance,

$$\text{bias}\left(\hat{\sigma}^2_{\text{ML}}\right) = -\frac{2}{n+1}\sigma^2_X,$$

which is slightly more of an underestimate than the ML estimator (when $n = 1$ the biases are equal) but the variance is less than both the UMVUE and ML estimators:

$$\mathbb{V}\left(\hat{\sigma}^2_{n+1}\right) = \frac{2(n-1)\sigma^4_X}{(n+1)^2}.$$

Furthermore, when $n > 1$, Hubert (1972) points out that mean squared error for the $n+1$ estimator,

$$\text{MSE}_{\text{ML}} = \frac{2\sigma^4_X}{n+1},$$

is the minimum mean squared error for all estimators of the population variance.

The ratio of the variances (i.e., the relative efficiency) of the UMVUE estimator and $n+1$ estimator is

$$\frac{\mathbb{V}\left(\hat{\sigma}^2_{\text{UMVUE}}\right)}{\mathbb{V}\left(\hat{\sigma}^2_{n+1}\right)} = \frac{(n+1)^2}{(n-1)^2};$$

as $n \to \infty$, the estimators are nearly equal in their bias and variance. In fact, it is worth noting that all three estimators are consistent; that is, for some $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\hat{\sigma}^2_{\text{UMVUE}} - \sigma^2_X| < \epsilon) = \lim_{n\to\infty} P(|\hat{\sigma}^2_{\text{ML}} - \sigma^2_X| < \epsilon) = \lim_{n\to\infty} P(|\hat{\sigma}^2_{n+1} - \sigma^2_X| < \epsilon) = 1.$$

By increasing the constant in the denominator of the estimator, the variance will continue to decrease while the bias continues to increase (in absolute terms)—an example of the variance-bias trade-off—but the minimum mean squared error is obtained using the $n+1$ estimator.

To illustrate, suppose $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 25)$. Simulating 10,000 independent samples of size $n = 10$ from this distribution gives the following estimates: $\widehat{\mathbb{E}}(\hat{\sigma}^2_{\text{UMVUE}}) = 24.99$ and $\widehat{\mathbb{V}}(\hat{\sigma}^2_{\text{UMVUE}}) = 134.97$; $\widehat{\mathbb{E}}(\hat{\sigma}^2_{n+1}) = 20.45$ and $\widehat{\mathbb{V}}(\hat{\sigma}^2_{n+1}) = 90.353$. The estimated mean squared

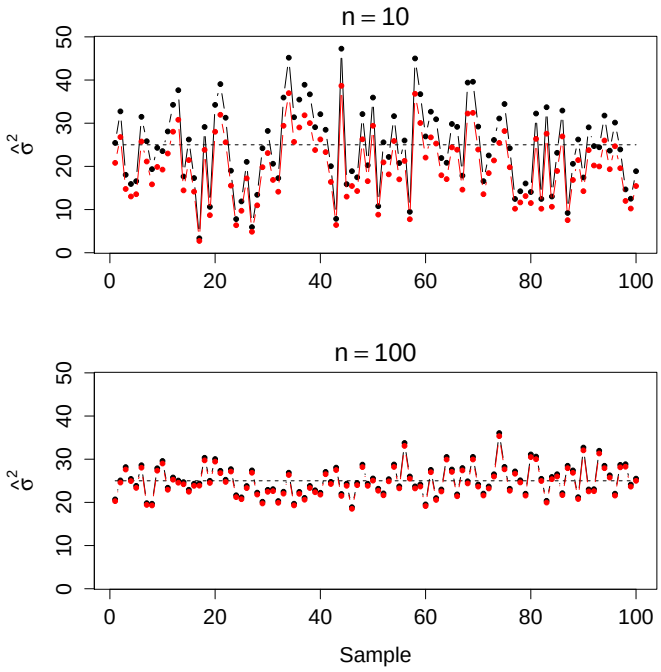

Figure 6.5: Variance of UMVUE and $n+1$ estimators of the population variance ($\sigma_X^2 = 25$) for sample size of $n = 10$ (top plot) and $n = 100$ (bottom plot).

errors are $\widehat{\text{MSE}}_{\text{UMVUE}} = 136.32$ and $\widehat{\text{MSE}}_{n+1} = 112.00$. Because $\sigma_X^2 = 25$, it is expected that $\mathbb{E}(\hat{\sigma}^2_{\text{UMVUE}}) = 25$, $\mathbb{V}(\hat{\sigma}^2_{\text{UMVUE}}) = \text{MSE}_{\text{UMVUE}} = 138.89$, $\mathbb{E}(\hat{\sigma}^2_{n+1}) = 20.45$, $\mathbb{V}(\hat{\sigma}^2_{n+1}) = 92.98$, and $\text{MSE}_{n+1} = 113.64$.

For a sample size of 100, the estimates are $\widehat{\mathbb{E}}(\hat{\sigma}^2_{\text{UMVUE}}) = 24.95$, $\widehat{\mathbb{V}}(\hat{\sigma}^2_{\text{UMVUE}}) = 12.28$, $\widehat{\mathbb{E}}(\hat{\sigma}^2_{n+1}) = 24.46$, $\widehat{\mathbb{V}}(\hat{\sigma}^2_{n+1}) = 11.80$, $\widehat{\text{MSE}}_{\text{UMVUE}} = 12.78$, and $\widehat{\text{MSE}}_{n+1} = 12.53$; the expected values are $\mathbb{E}(\hat{\sigma}^2_{\text{UMVUE}}) = 25$, $\mathbb{V}(\hat{\sigma}^2_{\text{UMVUE}}) = \text{MSE}_{\text{UMVUE}} = 12.63$, $\mathbb{E}(\hat{\sigma}^2_{n+1}) = 24.50$, $\mathbb{V}(\hat{\sigma}^2_{n+1}) = 12.13$, and $\widehat{\text{MSE}}_{n+1} = 12.38$. The top plot in Figure 6.5 gives the first 100 estimates of sample size $n = 10$ (the black lines represent the UMVUE estimates; the red lines, the $n+1$ estimates); similarly, the bottom plot displays the first 100 estimates of sample size $n = 100$. It is clear that for large enough n, the trade-off between bias and variance becomes negligible, but for small n, it is not.

6.3.2 Kelley True Score Estimator

In 1923, Truman Lee Kelley published *Statistical Method* where he discusses true score estimation (i.e., the estimation of the true ability of an individual) denoted here as X_T. Equation 168 on page 214 provides an estimator for the true score:

$$\hat{X}_T = \rho_X X + (1 - \rho_X)\mu_X, \tag{6.7}$$

where X is the observed score, μ_X is the overall mean of a subset of observations, and ρ_X is a reliability measure. Kelley's true score estimator regresses, or shrinks, the observed score toward the group mean; the less reliable the measure, the greater the shrinkage.

The classical true score model is

$$X = X_T + \varepsilon,$$

where $\mathbb{E}(\varepsilon|X_T) = 0$, $\mathbb{V}(\varepsilon|X_T) = \sigma^2_\varepsilon$, and $\text{Cov}(X_T, \varepsilon) = 0$ so that $\sigma^2_X = \sigma^2_{X_T} + \sigma^2_\varepsilon$. One choice for ρ_X is $\text{Corr}(X, X_T)^2$, where

$$\begin{aligned}\text{Corr}(X, X_T) &= \frac{\text{Cov}(X, X_T)}{\sigma_X \sigma_{X_T}} = \frac{\mathbb{E}(XX_T) - \mathbb{E}(X)\mathbb{E}(X_T)}{\sigma_X \sigma_{X_T}} \\ &= \frac{\mathbb{E}((X_T + \varepsilon)X_T) - \mathbb{E}(X_T)^2}{\sigma_X \sigma_{X_T}} = \frac{\mathbb{E}(X_T^2) - \mathbb{E}(X_T)^2}{\sigma_X \sigma_{X_T}} = \frac{\sigma_{X_T}}{\sigma_X};\end{aligned}$$

thus, $\rho_X = {}^{\sigma^2_{X_T}}/_{\sigma^2_X}$, which is the typical definition of the reliability coefficient.

Kelley's true score estimator can be directly related to the simple linear regression model where one would like to model X_T from X. The estimated simple linear regression of X_T regressed onto X is $\hat{X}_T = \hat{\beta}_0 + \hat{\beta}_1 X$ where $\hat{\beta}_1 = r_{XX_T}\,({}^{\hat{\sigma}_{X_T}}/_{\hat{\sigma}_X})$, $\beta_0 = \hat{\mu}_{X_T} - \hat{\beta}_1 \hat{\mu}_X$, and r_{XX_T} is the estimated correlation coefficient between X and X_T. Letting $r_{XX_T} = {}^{\hat{\sigma}_{X_T}}/_{\hat{\sigma}_X}$,

the estimated regression equation can be rewritten as

$$\begin{aligned}
\hat{X}_T &= \hat{\beta}_0 + \hat{\beta}_1 X \\
&= \hat{\mu}_X - r_{XX_T}\left(\frac{\hat{\sigma}_{X_T}}{\hat{\sigma}_X}\right)\hat{\mu}_X + r_{XX_T}\left(\frac{\hat{\sigma}_{X_T}}{\hat{\sigma}_X}\right)X \\
&= \left(1 - r^2_{XX_T}\right)\hat{\mu}_X + r^2_{XX_T}X \\
&= \hat{\rho}_X X + (1 - \hat{\rho}_X)\hat{\mu}_X.
\end{aligned}$$

The observed score, X, is an unbiased estimator of X_T; thus, Kelley's true score estimator, $\hat{X}_T$, is a biased estimator when $\rho_X \neq 1$ or $X \neq \mu_X$. The standard error of the observed score is $\sigma_X\sqrt{1-\rho_X} \geq \sigma_X\rho_X\sqrt{1-\rho_X}$, the standard error of Kelley's True Score estimator; thus, Kelley's True Score estimator is more efficient compared to the observed score (when $\rho_X \neq 0$). Said differently, Kelley's True Score estimator trades off increased squared bias for decreased variance.

6.3.3 James-Stein Estimator

Given a collection of n observations on p variables, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$, simultaneous estimation of their means is common (e.g., in an analysis of variance [ANOVA] setting). Suppose that $X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_j, \sigma^2_X)$ for $i = 1, \ldots, n$, $j = 1, \ldots, p$; here, one may want to estimate $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$.

An unbiased, sufficient, and consistent estimator of μ_j is the sample mean, $\bar{X}_j = \frac{1}{n}\sum_{i=1}^n X_{ij}$. However, given $p \geq 3$ variables, C. Stein (1956) showed that $\bar{\mathbf{X}} = (\bar{X}_1, \ldots, \bar{X}_p)$ is inadmissible; that is, $\mathcal{R}(\boldsymbol{\mu}, \bar{\mathbf{X}}) > \mathcal{R}(\boldsymbol{\mu}, \delta(\mathbf{X}))$ for some estimator $\delta(\mathbf{X})$ (for an empirical Bayes approach, see Efron & Morris, 1973). Such an estimator is provided by W. James and Stein (1961) and commonly referred to as the James-Stein estimator,

$$\bar{\mathbf{X}}_{JS} = \left(1 - \frac{(p-2)\sigma^2_{\bar{X}}}{\|\bar{\mathbf{X}}\|^2}\right)\bar{\mathbf{X}}, \tag{6.8}$$

where $\sigma^2_{\bar{X}} = \sigma^2_X/n$. Note that although $\mathcal{R}(\boldsymbol{\mu}, \bar{\mathbf{X}}) > \mathcal{R}(\boldsymbol{\mu}, \bar{\mathbf{X}}_{JS})$, the James-Stein estimator itself is not inadmissible.

The James-Stein estimator shrinks the estimator $\bar{\mathbf{X}}$ toward $\mathbf{0}$ when $(p-2)\sigma^2_{\bar{X}} < \|\bar{\mathbf{X}}\|^2$. Alternatively, one can shrink $\bar{\mathbf{X}}$ toward an arbitrarily-chosen vector, say $\boldsymbol{\mu}_{\bar{\mathbf{X}}}$, where each entry is equal to the overall mean $\frac{1}{p}\sum_{j=1}^{p} \bar{X}_j$. The James-Stein estimator is then

$$\bar{\mathbf{X}}_{JS} = \left(1 - \frac{(p-2)\sigma^2_{\bar{X}}}{\|\bar{\mathbf{X}} - \boldsymbol{\mu}_{\bar{\mathbf{X}}}\|^2}\right)(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\bar{\mathbf{X}}}) + \boldsymbol{\mu}_{\bar{\mathbf{X}}}.$$

This can be rewritten as

$$\begin{aligned}\bar{\mathbf{X}}_{JS} &= \left(1 - \frac{(p-2)\sigma^2_{\bar{X}}}{\|\bar{\mathbf{X}} - \boldsymbol{\mu}_{\bar{\mathbf{X}}}\|^2}\right)\bar{\mathbf{X}} + \left(1 - 1 - \frac{(p-2)\sigma^2_{\bar{X}}}{\|\bar{\mathbf{X}} - \boldsymbol{\mu}_{\bar{\mathbf{X}}}\|^2}\right)\boldsymbol{\mu}_{\bar{\mathbf{X}}} \\ &= \rho_{\bar{\mathbf{X}}}\bar{\mathbf{X}} + (1 - \rho_{\bar{\mathbf{X}}})\,\hat{\boldsymbol{\mu}}_{\bar{\mathbf{X}}},\end{aligned}$$

where

$$\rho_{\bar{\mathbf{X}}} = \left(1 - \frac{(p-2)\sigma^2_{\bar{X}}}{\|\bar{\mathbf{X}} - \boldsymbol{\mu}_{\bar{\mathbf{X}}}\|^2}\right).$$

When $(p-2)\sigma^2_{\bar{X}} < \|\bar{\mathbf{X}} - \boldsymbol{\mu}_{\bar{\mathbf{X}}}\|^2$, the estimator shrinks $\bar{\mathbf{X}}$ toward $\boldsymbol{\mu}_{\bar{\mathbf{X}}}$.

6.3.4 Example

As a demonstration, suppose N patients are evaluated using eleven distinct violence prediction methods and all N of these patients will commit an act of violence in the future. Let $X_i \overset{\text{iid}}{\sim} \mathcal{B}\text{ernoulli}(p_i)$ where $p_i = .60, .62, \ldots, .80$ and $i = 1, \ldots, 11$; that is, p_i represents the true sensitivity for each of the eleven distinct prediction devices. Kelley's True Score (TS) estimator is used to estimate the "true" sensitivity of the eleven violence prediction methods and this is compared to the mean squared error of the observed sensitivity, $\hat{p}^{(i)}_{B|A} = \frac{1}{N}\sum_{j=1}^{N} X_{ij}$. The reliability coefficient is estimated using Hoyt's method (Hoyt, 1941), where in Hoyt's terminology, the prediction method represents the "student" and the patients are

the "items." The simulation code can be found in Appendix C.

For $N = 20$, the estimated reliability coefficient is .610. The observed proportions are $\hat{\mathbf{p}}_{B|A} = (.50,.65,.40,.75,.90,.85,.80,.60,.75,.75,.85)^T$; the estimates using Kelley's TS estimator are $\hat{\mathbf{p}}_{\text{TS}} = (.582,.673,.521,.734,.825,.795,.765,.643,.734,.734,.795)^T$. Eight of the eleven Kelley TS estimates are closer to the true sensitivity than the observed sensitivity. In addition, the mean squared error for Kelley's TS estimator ($\widehat{\text{MSE}} = 0.006$) is about a third of the size when using the observed sensitivities ($\widehat{\text{MSE}} = 0.016$).

Increasing the number of observations reduces, but does not diminish, the benefit of using Kelley's TS estimator. For $N = 100$, six of Kelley's TS estimates are closer to the true value; the mean squared error for Kelley's TS estimator is 0.0011 compared to 0.0012 using the observed sensitivities. The advantage of Kelley's True Score estimator is still present, but it has decreased with a larger number of observations.

6.3.5 Ridge Regression

Regularization imposes a penalty term in the fitting of a regression model as a way of preventing overfitting of the model. There are several methods of regularization, but only *ridge regression* is discussed here (Hoerl & Kennard, 1970) based on Tikhonov regularization (Tikhonov, 1943, 1963). Two other popular regularization tools are the *Lasso* (least absolute shrinkage and selection operator; Tibshirani, 1996) and *elastic net* (a weighted combination of ridge regression and the Lasso; Zou & Hastie, 2005). Both methods are similar to ridge regression and are shrinkage estimators; unlike ridge regression they also perform variable selection and should be preferred to techniques such as stepwise regression which is prone to overfitting (e.g., see Babyak, 2004).

Consider the $n \times p$ dataset, $\mathbf{X}$, containing p mean-centered predictor variables for n observations, and a mean-centered response variable, $\mathbf{y}$, of size $n \times 1$ (the variables are assumed to be mean-centered for simplicity of notation, but they need not be). In addition, $\mathbf{X}$ should be standardized so that $\mathbb{V}(\mathbf{X}_j) = \sigma_X^2$ for $j = 1, \ldots, p$ (typically $\sigma_X^2 = 1$) because,

unlike the least squares estimator for regression, the ridge regression estimator is not scale equivariant (i.e., the ridge estimator is affected by the scale of the variables). In least squares regression, one minimizes the squared loss function to obtain the unique (when $\mathbf{X}$ is of full column rank, which is assumed throughout) least squares solution for the regression coefficients; that is,

$$\hat{\boldsymbol{\beta}}_{\text{LS}} = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2, \tag{6.9}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the $p \times 1$ vector containing the estimates for the linear regression function $f(x) = \beta_1 x_1 + \cdots + \beta_p x_p$.

Ridge regression minimizes the slightly different objective function,

$$\hat{\boldsymbol{\beta}}^R = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2, \tag{6.10}$$

where $\lambda \geq 0$ is called a *tuning parameter* (cross-validation can be used to determine an appropriate value of λ). Note in the trivial case where $\lambda = 0$, Equation (6.10) is equivalent to Equation (6.9), the least squares solution. Increasing λ tends to decrease the estimates of $\boldsymbol{\beta}$; as $\lambda \to \infty$, $\hat{\boldsymbol{\beta}}^R \to \mathbf{0}_p$, where $\mathbf{0}_p$ is a $p \times 1$ vector of 0's. Thus, ridge regression is another type of shrinkage estimator and the term $\lambda\|\boldsymbol{\beta}\|^2$ can be thought of as a shrinkage penalty.

The Gauss-Markov theorem states that the least squares (LS) estimator of $\boldsymbol{\beta}$ (i.e., $\hat{\boldsymbol{\beta}}_{\text{LS}}$) is the best linear unbiased estimator of $\boldsymbol{\beta}$; that is, of all unbiased estimators, $\hat{\boldsymbol{\beta}}_{\text{LS}}$ has minimum variance. Therefore, if one were to choose an unbiased estimator for $\boldsymbol{\beta}$, the least squares estimator is the obvious choice; however, one may desire an estimator that is more efficient than the LS estimator (i.e, one that has smaller variance). Clearly, this estimator is biased because there does not exist an unbiased estimator that is more efficient than the least squares one; thus, there again is a trade-off between variance and bias. In fact, increasing λ decreases the flexibility of the ridge regression model and this typically leads to decreased

variance and increased bias.

Equation (6.10) can be considered a constrained optimization problem; that is, minimizing Equation (6.10) for a given λ is equivalent to minimizing

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \qquad \text{subject to} \quad \|\boldsymbol{\beta}\|^2 \leq s,$$

where s is inversely proportional to λ (i.e., a larger s corresponds to a smaller λ). Geometrically, $\|\boldsymbol{\beta}\|^2 \leq s$ represents a hypersphere in R^p.

To solve for $\hat{\boldsymbol{\beta}}^R$, Equation (6.10) can be rewritten as

$$\hat{\boldsymbol{\beta}}^R = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \|\mathbf{0}_p - \sqrt{\lambda}\mathbf{I}_p\boldsymbol{\beta}\|^2,$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix. Letting

$$\mathbf{y}_\lambda = \begin{pmatrix} \mathbf{y} \\ \mathbf{0}_p \end{pmatrix}$$

and

$$\mathbf{X}_\lambda = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I}_p \end{pmatrix},$$

Equation (6.10) can be re-expressed as

$$\|\mathbf{y}_\lambda - \mathbf{X}_\lambda\boldsymbol{\beta}\|^2,$$

where

$$\hat{\boldsymbol{\beta}}^R = (\mathbf{X}_\lambda^T\mathbf{X}_\lambda)^{-1}\mathbf{X}_\lambda^T\mathbf{y}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y} \tag{6.11}$$

is the solution. As one might expect, the ridge regression estimator is a biased estimator of $\boldsymbol{\beta}$ when $\lambda > 0$:

$$\mathbb{E}\left(\hat{\boldsymbol{\beta}}^R\right) = \left(\mathbf{I}_p + \lambda(\mathbf{X}^T\mathbf{X})^{-1}\right)^{-1}\boldsymbol{\beta}.$$

When the data are orthonormal so that $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$,

$$\hat{\boldsymbol{\beta}}^R = \left(\frac{1}{1+\lambda}\right)\hat{\boldsymbol{\beta}}_{\text{LS}}.$$

As a simple example, suppose that $n = p$ and $\mathbf{X} = \mathbf{I}_p$; Equation (6.11) reduces to $(\mathbf{I}_p + \lambda\mathbf{I}_p)^{-1}\mathbf{y}$, so that for $j = 1, \ldots, p$,

$$\hat{\beta}_j^R = \frac{y_j}{1+\lambda}.$$

Because $\bar{y} = 0$,

$$\hat{\beta}_j^R = \frac{y_j}{1+\lambda} = \frac{y_j - \bar{y}}{1+\lambda} + \bar{y} = \frac{1}{1+\lambda}y_j + \left(1 - \frac{1}{1+\lambda}\right)\bar{y}.$$

For this simple situation, the ridge regression estimator is a special case of Kelley's True Score estimator where $\rho = {}^1\!/_{(1+\lambda)}$.

Ridge Regression Estimator and the Singular Value Decomposition.

Let $\mathbf{X} = \mathbf{UDV}^T$ be the singular value decomposition (SVD) of $\mathbf{X}$, where $\mathbf{U}$ is an $n \times p$ matrix containing the left singular vectors ($\mathbf{U}^T\mathbf{U} = \mathbf{I}$), $\mathbf{D}$ is a $p \times p$ diagonal matrix containing the singular values $\theta_1 \geq \theta_2 \cdot \cdots \geq \theta_p \geq 0$ (because the variables have common variance, if the variables are also uncorrelated $\theta_1 = \theta_2 = \cdots = \theta_p \geq 0$), and $\mathbf{V}$ is a $p \times p$ orthogonal matrix containing the right singular vectors. Using the SVD of $\mathbf{X}$, the estimator $\hat{\boldsymbol{\beta}}^R$ can be rewritten as

$$\begin{aligned}\hat{\boldsymbol{\beta}}^R &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y} \\ &= (\mathbf{VDU}^T\mathbf{UDV}^T + \lambda\mathbf{I}_p)^{-1}\mathbf{VDU}^T\mathbf{y}\end{aligned}$$

$$
\begin{aligned}
&= (\mathbf{V}\mathbf{D}^2\mathbf{V}^T + \lambda\mathbf{I}_p)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} \\
&= \left(\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I}_p)\mathbf{V}^T\right)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} \\
&= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y} \\
&= \mathbf{V}\mathbf{D}_R\mathbf{U}^T\mathbf{y},
\end{aligned}
$$

where $\mathbf{D}_R$ is a diagonal matrix such that $\{\mathbf{D}_R\}_{jj} = \theta_j/(\theta_j^2+\lambda)$. Thus, $\mathbf{U}\mathbf{D}_R\mathbf{V}^T$ is the singular value decomposition of $\left(\hat{\boldsymbol{\beta}}^R\right)^T = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}$.

The least squares solution can be similarly decomposed as

$$\hat{\boldsymbol{\beta}}_{\text{LS}} = \mathbf{V}\mathbf{D}_{\text{LS}}\mathbf{U}^T\mathbf{y},$$

where $\{\mathbf{D}_{\text{LS}}\}_{jj} = 1/\theta_j$.

The estimated values of $\mathbf{y}$ are $\hat{\mathbf{y}}_\lambda = \mathbf{X}\hat{\boldsymbol{\beta}}^R = \mathbf{H}_\lambda\mathbf{y}$ where $\mathbf{H}_\lambda = \mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^T$ is a symmetric matrix that is analogous to the so-called "hat matrix" in least squares regression (but note this matrix is not idempotent so it is not a projection matrix). The estimated response can be expressed in terms of the SVD of $\mathbf{X}$,

$$\hat{\mathbf{y}}_\lambda = \mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{V}\mathbf{D}_R\mathbf{U}^T\mathbf{y} = \mathbf{U}\mathbf{D}_\lambda\mathbf{U}^T\mathbf{y},$$

where $\{\mathbf{D}_\lambda\}_{jj} = \theta_j^2/(\theta_j^2+\lambda)$; thus, $\mathbf{U}\mathbf{D}_\lambda\mathbf{U}^T$ is the singular value decomposition of $\mathbf{H}_\lambda$.

The square of the singular values of $\mathbf{X}$ are the eigenvalues of $\mathbf{X}^T\mathbf{X}$ so $1/\theta_j^2$ is the $(p-j+1)$th eigenvalue of $(\mathbf{X}^T\mathbf{X})^{-1}$. The eigenvalues and eigenvectors, $\mathbf{v}_j$, are directly related to principal components analysis where $\mathbf{v}_j$ is the jth principal component of $(\mathbf{X}^T\mathbf{X})$ and σ_j^2 is the variance associated with it. Furthermore,

$$
\begin{aligned}
\hat{\mathbf{y}}_\lambda &= \mathbf{X}\boldsymbol{\beta}^R \\
&= \mathbf{U}\mathbf{D}_\lambda\mathbf{U}^T\mathbf{y}
\end{aligned}
$$

$$= \sum_{j=1}^{p} d_{\lambda j} \mathbf{u}_j \mathbf{u}_j^T \mathbf{y}.$$

Thus, $\mathbf{y}$ is projected to the singular vector space for $\mathbf{H}_\lambda$ ($\mathbf{U}\mathbf{U}^T$ is a projection matrix), and then shrunken by the singular values of $\mathbf{H}_\lambda$.

Comparison of Ridge Regression with Least Squares.

Suppose that

$$\mathbf{y} = f(\mathbf{X}) + \boldsymbol{\varepsilon},$$

where $\mathbb{E}(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}_n$ and $\mathbb{V}(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I}_n$. Given $(\mathbf{y}, \mathbf{X})$, $f(\mathbf{X})$ is estimated as $\hat{\mathbf{y}} \equiv \hat{f}(\mathbf{X})$. Let $(\mathbf{y}_*, \mathbf{X}_*)$ be a set of n' new observations (e.g., the testing sample) and $\mathbf{y}_* = f(\mathbf{X}_*) + \boldsymbol{\varepsilon}_*$ where $\mathbb{E}(\boldsymbol{\varepsilon}_*|\mathbf{X}_*) = \mathbf{0}_{n'}$ and $\mathbb{V}(\boldsymbol{\varepsilon}_*|\mathbf{X}_*) = \sigma_{\boldsymbol{\varepsilon}}^2 \mathbf{I}_{n'}$ (note that $f(\cdot)$ is the same). Here, $\mathbf{y}_*$ and $\mathbf{y}$ are assumed independent. Let $\hat{\mathbf{y}}_{\text{LS}} = \mathbf{X}_* \hat{\boldsymbol{\beta}}_{\text{LS}} = \mathbf{X}_* (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ represent the least squares estimates for the new observations. The measure of interest is the expected error for the new observations:

$$\mathbb{E}\left(\|\mathbf{y}_* - \hat{\mathbf{y}}_{\text{LS}}\|^2 | \mathbf{X}, \mathbf{X}_* \right) = \| \mathbb{E}\left(\mathbf{y}_* - \hat{\mathbf{y}}_{\text{LS}} | \mathbf{X}, \mathbf{X}_* \right) \|^2 + \text{tr}\left(\mathbb{V}(\mathbf{y}_* - \hat{\mathbf{y}}_{\text{LS}}) | \mathbf{X}, \mathbf{X}_* \right).$$

Because $\hat{\boldsymbol{\beta}}_{\text{LS}}$ is an unbiased estimator of $\boldsymbol{\beta}$, $\mathbb{E}\left(\mathbf{y}_* - \hat{\mathbf{y}}_{\text{LS}} \right) = \mathbf{0}_{n'}$.

Note that $\hat{\mathbf{y}}_{\text{LS}}$ depends only on $\mathbf{y}$, and $\mathbf{y}$ and $\mathbf{y}_*$ are independent, so $\mathbf{y}_*$ and $\hat{\mathbf{y}}_{\text{LS}}$ are also independent. Thus,

$$\begin{aligned}
\mathbb{V}(\mathbf{y}_* - \hat{\mathbf{y}}_{\text{LS}} | \mathbf{X}, \mathbf{X}_*) &= \mathbb{V}(\mathbf{y}_* | \mathbf{X}_*) + \mathbb{V}(\hat{\mathbf{y}}_{\text{LS}} | \mathbf{X}) \\
&= \sigma_{\varepsilon}^2 \mathbf{I}_{n'} + \mathbf{X}_* (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{V}(\mathbf{y}|\mathbf{X}) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_*^T \\
&= \sigma_{\varepsilon}^2 \mathbf{I}_{n'} + \sigma_{\varepsilon}^2 \mathbf{X}_* (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_*^T.
\end{aligned}$$

The expected error reduces to

$$\begin{aligned}\mathbb{E}\left(\|\mathbf{y}_* - \hat{\mathbf{y}}_{\text{LS}}\|^2|\mathbf{X}, \mathbf{X}_*\right) &= \sigma_\varepsilon^2\left(\text{tr}(\mathbf{I}_{n'}) + \text{tr}\left(\mathbf{X}_*(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_*^T\right)\right) \\ &= n'\sigma_\varepsilon^2 + \sigma_\varepsilon^2\text{tr}\left(\mathbf{X}_*^T\mathbf{X}_*(\mathbf{X}^T\mathbf{X})^{-1}\right).\end{aligned}$$

Because $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, $(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T$ and letting the singular value decomposition of $\mathbf{X}_*$ be $\mathbf{U}_*\mathbf{D}_*\mathbf{V}_*^T$, where $\mathbf{U}_*$ is $n' \times p$ and $\mathbf{D}_*$ and $\mathbf{V}_*$ are $p \times p$, gives $\mathbf{X}_*^T\mathbf{X}_* = \mathbf{V}_*\mathbf{D}_*^2\mathbf{V}_*^T$; thus,

$$\text{tr}\left(\mathbf{X}_*^T\mathbf{X}_*(\mathbf{X}^T\mathbf{X})^{-1}\right) = \text{tr}\left(\mathbf{V}_*\mathbf{D}_*^2\mathbf{V}_*^T\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T\right).$$

The above equation can be simplified further:

$$\text{tr}\left(\mathbf{V}_*\mathbf{D}_*^2\mathbf{V}_*^T\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T\right) = \sum_{k=1}^{p}\sum_{j=1}^{p} v_{*kj}^2 v_{kj}^2\left(\frac{\theta_{*j}^2}{\theta_j^2}\right) = \sum_{k=1}^{p}\sum_{j=1}^{p} s_{jk}^2$$

where $s_{jk} = v_{*kj}v_{kj}\left(\theta_{*j}/\theta_j\right)$, $j, k = 1, \ldots, p$.

Dividing the expected error by n' gives the average expected error,

$$\frac{1}{n'}\mathbb{E}\left(\|\mathbf{y}_* - \hat{\mathbf{y}}_{\text{LS}}\|^2|\mathbf{X}, \mathbf{X}_*\right) = \left(\frac{\sigma_\varepsilon^2}{n'}\right)\sum_{k=1}^{p}\sum_{j=1}^{p} s_{jk}^2 + \sigma_\varepsilon^2,$$

where σ_ε^2 is the irreducible error and $\left(\sigma_\varepsilon^2/n'\right)\sum_{k=1}^{p}\sum_{j=1}^{p} s_{jk}^2$ is the variance associated with the model (this is equal to the reducible error because the estimator is unbiased).

Suppose $\mathbf{X}_* = \mathbf{X}$ so that $\mathbb{V}(\hat{\mathbf{y}}_{\text{LS}}|\mathbf{X}) = \sigma_\varepsilon^2\mathbf{H}$ and $\mathbb{E}\left(\|\mathbf{y}_* - \hat{\mathbf{y}}_{\text{LS}}\|^2|\mathbf{X}, \mathbf{X}_*\right) = n'\sigma_\varepsilon^2 + p\sigma_\varepsilon^2$ so the reducible error is $\left(p/n'\right)\sigma_\varepsilon^2$. Note that $\mathbf{y}_* \neq \mathbf{y}$; rather new values on the response are considered corresponding to the observations on the predictor variables used to construct the model. Because the true function, $f(\mathbf{X})$, depends on the p predictor variables, the reducible error is reduced when new observations are closer to 0 (the means of the predictor variables); that is, by reducing $\|\mathbf{x}_{*i}\|^2$. When $\mathbf{X}_* = \mathbf{X}$, reduction in the variance is achieved by increasing the sample size.

For the ridge regression estimator, the estimates for the new observations are $\hat{\mathbf{y}}_\lambda = \mathbf{X}_*\hat{\boldsymbol{\beta}}^R = \mathbf{X}_*(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}$. Similar to the least squares solution, the expected error for the ridge regression estimator can be split into two parts:

$$\mathbb{E}\left(\|\mathbf{y}_* - \hat{\mathbf{y}}_\lambda\|^2|\mathbf{X}, \mathbf{X}_*\right) = \|\mathbb{E}(\mathbf{y}_*|\mathbf{X}_*) - \mathbb{E}(\hat{\mathbf{y}}_\lambda|\mathbf{X})\|^2 + \operatorname{tr}\left(\mathbb{V}(\mathbf{y}_* - \hat{\mathbf{y}}_\lambda|\mathbf{X}, \mathbf{X}_*)\right)$$

The first part reduces to

$$\begin{aligned}\|\mathbb{E}(\mathbf{y}_*|\mathbf{X}_*) - \mathbb{E}(\hat{\mathbf{y}}_\lambda|\mathbf{X})\|^2 &= \|\mathbf{X}_*\boldsymbol{\beta} - \mathbf{X}_*(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbb{E}(\mathbf{y})\|^2 \\ &= \|\mathbf{X}_*\left(\mathbf{I}_p - (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta}\|^2.\end{aligned}$$

When $\lambda = 0$ the above equation is zero; this corresponds to the least squares solution and is consistent with the result found earlier. Because $\mathbf{y}_*$ and $\mathbf{y}$ are independent, the covariance term in the second part of the expected error becomes

$$\begin{aligned}\mathbb{V}(\mathbf{y}_* - \hat{\mathbf{y}}_\lambda|\mathbf{X}, \mathbf{X}_*) &= \mathbb{V}(\mathbf{y}_*|\mathbf{X}_*) + \mathbb{V}(\hat{\mathbf{y}}_\lambda|\mathbf{X}) \\ &= \sigma_\varepsilon^2\mathbf{I}_{n'} + \mathbf{X}_*(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbb{V}(\mathbf{y}|\mathbf{X})\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}_*^T \\ &= \sigma_\varepsilon^2\left(\mathbf{I}_{n'} + \mathbf{X}_*(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}_*^T\right).\end{aligned}$$

The trace of the above expression is

$$\begin{aligned}\operatorname{tr}\left(\mathbb{V}(\mathbf{y}_* - \hat{\mathbf{y}}_\lambda|\mathbf{X}, \mathbf{X}_*)\right) &= \operatorname{tr}\left(\sigma_\varepsilon^2\left(\mathbf{I}_{n'} + \mathbf{X}_*(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}_*^T\right)\right) \\ &= n'\sigma_\varepsilon^2 + \sigma_\varepsilon^2\operatorname{tr}\left(\mathbf{X}_*^T\mathbf{X}_*(\mathbf{I}_p + \lambda(\mathbf{X}^T\mathbf{X})^{-1})^{-1}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\right) \\ &= n'\sigma_\varepsilon^2 + \sigma_\varepsilon^2\operatorname{tr}\left(\mathbf{V}_*\mathbf{D}_*^2\mathbf{V}_*^T\mathbf{V}(\mathbf{I}_p + \lambda\mathbf{D}^{-2})^{-1}\mathbf{V}^T\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{V}^T\right) \\ &= n'\sigma_\varepsilon^2 + \sigma_\varepsilon^2\operatorname{tr}\left(\mathbf{X}_*^T\mathbf{X}_*\mathbf{V}\mathbf{D}^2(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-2}\mathbf{V}^T\right)\end{aligned}$$

The second part can be rewritten as

$$\sum_{k=1}^{p}\sum_{j=1}^{p} v_{*kj}^2 v_{kj}^2 \frac{\theta_j^2\theta_{*j}^2}{(\theta_j^2+\lambda)^2} = \sum_{k=1}^{p}\sum_{j=1}^{p} s_{\lambda jk}^2,$$

where $s_{\lambda jk} = v_{*kj}v_{kj}\left(\theta_j\theta_{*j}/(\theta_j^2+\lambda)\right)$, $j,k = 1,\ldots,p$.

Putting this together and averaging over the n' observations gives the average estimated error:

$$\frac{1}{n'}\mathbb{E}\left(\|\mathbf{y}_* - \hat{\mathbf{y}}_\lambda\|^2|\mathbf{X},\mathbf{X}_*\right) = \frac{1}{n'}\|\mathbf{X}_*\left(\mathbf{I}_p - (\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta}\|^2 + \left(\frac{\sigma_\varepsilon^2}{n'}\right)\sum_{k=1}^{p}\sum_{j=1}^{p} s_{\lambda jk}^2 + \sigma_\varepsilon^2.$$

The difference in the squared bias between the least squares and ridge estimators is the bias of the ridge estimator,

$$\|\mathbf{X}_*\left(\mathbf{I}_p - (\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta}\|^2.$$

Using the SVD of $\mathbf{X}$ leads to the following:

$$\begin{aligned}
\mathbf{X}_*\left(\mathbf{I}_p - (\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta} &= \mathbf{X}_*\left(\mathbf{I}_p - (\mathbf{V}\mathbf{D}^2\mathbf{V}^T+\lambda\mathbf{I}_p)^{-1}\mathbf{V}\mathbf{D}^2\mathbf{V}^T\right)\boldsymbol{\beta} \\
&= \mathbf{X}_*\left(\mathbf{I}_p - (\mathbf{V}(\mathbf{D}^2+\lambda\mathbf{I}_p)\mathbf{V}^T)^{-1}\mathbf{V}\mathbf{D}^2\mathbf{V}^T\right)\boldsymbol{\beta} \\
&= \mathbf{X}_*\left(\mathbf{I}_p - \mathbf{V}(\mathbf{D}^2+\lambda\mathbf{I}_p)^{-1}\mathbf{D}^2\mathbf{V}^T\right)\boldsymbol{\beta} \\
&= \mathbf{X}_*\left(\mathbf{I}_p - \mathbf{V}\mathbf{D}_\lambda\mathbf{V}^T\right)\boldsymbol{\beta}.
\end{aligned}$$

When $\lambda = 0$, $\mathbf{I}_p - \mathbf{V}\mathbf{D}_\lambda\mathbf{V}^T = \mathbf{0}_{p\times p}$ and as $\lambda\to\infty$, $\mathbf{I}_p - \mathbf{V}\mathbf{D}_\lambda\mathbf{V}^T \to \mathbf{I}_p$ so the bias depends on the size of λ and $\mathbf{X}_*$. The squared bias for the for the ridge estimator is

$$\begin{aligned}
\|\mathbf{X}_*\left(\mathbf{I}_p - \mathbf{V}\mathbf{D}_\lambda\mathbf{V}^T\right)\boldsymbol{\beta}\|^2 &= (\mathbf{X}_*\left(\mathbf{I}_p - \mathbf{V}\mathbf{D}_\lambda\mathbf{V}^T\right)\boldsymbol{\beta})^T(\mathbf{X}_*\left(\mathbf{I}_p - \mathbf{V}\mathbf{D}_\lambda\mathbf{V}^T\right)\boldsymbol{\beta}) \\
&= \boldsymbol{\beta}^T\left(\mathbf{I}_p - \mathbf{V}\mathbf{D}_\lambda\mathbf{V}^T\right)\mathbf{X}_*^T\mathbf{X}_*\left(\mathbf{I}_p - \mathbf{V}\mathbf{D}_\lambda\mathbf{V}^T\right)\boldsymbol{\beta}
\end{aligned}$$

Clearly the bias will always be larger for the ridge estimator ($\lambda > 0$) than for least

squares ($\lambda = 0$), but what about the variance? The difference in variance between the estimates

$$\sigma_\varepsilon^2) \sum_{k=1}^{p} \sum_{j=1}^{p} \left(s_{jk}^2 - s_{\lambda jk}^2 \right);$$

this leads to the following:

$$\begin{aligned}
\left(s_{jk}^2 - s_{\lambda jk}^2 \right) &= v_{*kj}^2 v_{kj}^2 \left(\frac{\theta_{*j}^2}{\theta_j^2} - \frac{\theta_j^2 \theta_{*j}^2}{(\theta_j^2 + \lambda)^2} \right) \\
&= v_{*kj}^2 v_{kj}^2 \left(\frac{\theta_{*j}^2 (\theta_j^2 + \lambda)^2 - \theta_j^2 (\theta_j^2 \theta_{*j}^2)}{\theta_j^2 (\theta_j^2 + \lambda)^2} \right) \\
&= v_{*kj}^2 v_{kj}^2 \left(\frac{\theta_{*j}^2 (\theta_j^4 + 2\theta_j^2 \lambda + \lambda^2) - \theta_j^4 \theta_{*j}^2}{\theta_j^2 (\theta_j^2 + \lambda)^2} \right) \\
&= v_{*kj}^2 v_{kj}^2 \left(\frac{2\theta_{*j}^2 \theta_j^2 \lambda + \theta_{*j}^2 \lambda^2}{\theta_j^2 (\theta_j^2 + \lambda)^2} \right).
\end{aligned}$$

Because $v_{*kj}^2 v_{kj}^2 \geq 0$, $\theta_j^2 \geq 0$, $\theta_{*j}^2 \geq 0$, and $\lambda \geq 0$ for all $j, k = 1, \ldots, p$, the above expression is always nonnegative; in other words, the variance for the least squares estimator is always greater than (or in the trivial cases, equal to) the ridge estimator. Davis-Stober, Dana, and Budescu (2010) provide similar results for a set of constrained estimators for the regression coefficient that include the ridge estimator. Similar to the estimator for the population variance, the variance of the ridge estimator for $\boldsymbol{\beta}$ can be made infinitesimally small by increasing λ, but also similarly, this is at the cost of increasing the squared bias.

Taking the difference of the expected errors between the least squares estimator and the ridge regression estimator gives

$$\begin{aligned}
&\mathbb{E}\left(\|\mathbf{y}_* - \hat{\mathbf{y}}_{\text{LS}}\|^2 | \mathbf{X}, \mathbf{X}_* \right) - \mathbb{E}\left(\|\mathbf{y}_* - \hat{\mathbf{y}}_\lambda\|^2 | \mathbf{X}, \mathbf{X}_* \right) \\
&\qquad = \sigma_\varepsilon^2 \sum_{k=1}^{p} \sum_{j=1}^{p} \left(s_{jk}^2 - s_{\lambda jk}^2 \right) - \|\mathbf{X}_* \left(\mathbf{I}_p - (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{X} \right) \boldsymbol{\beta}\|^2
\end{aligned}$$

Thus, the mean squared error for the ridge estimator will be less than that for the least

squares estimator when

$$\sum_{k=1}^{p}\sum_{j=1}^{p} v_{*kj}^2 v_{kj}^2 \left(\frac{2\theta_{*j}^2\theta_j^2\lambda + \theta_{*j}^2\lambda^2}{\theta_j^2(\theta_j^2+\lambda)^2}\right)\sigma_\varepsilon^2 > \|\mathbf{X}_*\left(\mathbf{I}_p - (\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta}\|^2$$

In the case when the data are orthonormal so that $\mathbf{X}$ and $\mathbf{X}_*$ are orthogonal the variance of the ridge estimator simplifies to

$$n'\sigma_\varepsilon^2 + p\sigma_\varepsilon^2\left(\frac{1}{1+\lambda}\right)^2$$

and the variance of the least squares estimator is

$$(n'+p)\sigma_\varepsilon^2;$$

clearly the variance is less for the ridge estimator than the least squares one. The bias of the ridge estimator simplifies to

$$\left(\frac{\lambda}{1+\lambda}\right)^2\|\boldsymbol{\beta}\|^2.$$

Thus, the mean squared error is for the ridge estimator is

$$\mathbb{E}\left(\|\mathbf{y}_* - \hat{\mathbf{y}}_\lambda\|^2|\mathbf{X},\mathbf{X}_*\right) = \left(\frac{\lambda}{1+\lambda}\right)^2\|\boldsymbol{\beta}\|^2 + p\sigma_\varepsilon^2\left(\frac{1}{1+\lambda}\right)^2 + n'\sigma_\varepsilon^2,$$

Taking the derivative with respect to λ,

$$\frac{\partial}{\partial\lambda}\left(\mathbb{E}\left(\|\mathbf{y}_* - \hat{\mathbf{y}}_\lambda\|^2|\mathbf{X},\mathbf{X}_*\right)\right) = \left(\frac{2\lambda}{(1+\lambda)^3}\right)^2\|\boldsymbol{\beta}\|^2 - \left(\frac{2}{(1+\lambda)^3}\right)p\sigma_\varepsilon^2;$$

setting this to zero and solving for λ gives

$$\lambda = \frac{p\sigma_\varepsilon^2}{\|\boldsymbol{\beta}\|^2},$$

The second derivative evaluated at the above critical value is

$$\frac{2p\sigma_\varepsilon^2 + 2\|\boldsymbol{\beta}\|^2}{\left(1 - \frac{p\sigma_\varepsilon^2}{\|\boldsymbol{\beta}\|^2}\right)^4},$$

which is always positive; thus, the critical value is a global minimum meaning that the mean squared error is minimized when $\lambda = \frac{p\sigma_\varepsilon^2}{\|\boldsymbol{\beta}\|^2}$. Assuming $p \geq 1$, σ_ε^2, and $\boldsymbol{\beta} \neq \mathbf{0}$, the critical value is always greater than zero meaning the mean squared error is never minimized for the least squares solution. In practice σ_ε^2 and $\boldsymbol{\beta}$ are unknown but, as mentioned earlier, cross-validation can assist in determining λ.

The difference between the mean squared error for the least squares and ridge estimators when $\mathbf{X}$ is orthogonal is

$$\begin{aligned}\mathbb{E}\left(\|\mathbf{y}_* - \hat{\mathbf{y}}_{\text{LS}}\|^2 | \mathbf{X}, \mathbf{X}_*\right) - \mathbb{E}\left(\|\mathbf{y}_* - \hat{\mathbf{y}}_\lambda\|^2 | \mathbf{X}, \mathbf{X}_*\right) &= (n' + p)\sigma_\varepsilon^2 - n'\sigma_\varepsilon^2 - p\sigma_\varepsilon^2 \left(\frac{1}{1+\lambda}\right)^2 - \left(\frac{\lambda}{1+\lambda}\right)^2 \|\boldsymbol{\beta}\|^2 \\ &= p\sigma_\varepsilon^2 \left(1 - \left(\frac{1}{1+\lambda}\right)^2\right) - \left(\frac{\lambda}{1+\lambda}\right)^2 \|\boldsymbol{\beta}\|^2.\end{aligned}$$

The ridge estimator has a smaller mean squared error than the least squares one when

$$p\sigma_\varepsilon^2 \left(1 - \left(\frac{1}{1+\lambda}\right)^2\right) > \left(\frac{\lambda}{1+\lambda}\right)^2 \|\boldsymbol{\beta}\|^2.$$

This can be reduced to

$$\frac{2+\lambda}{\lambda} > \frac{\|\boldsymbol{\beta}\|^2}{p\sigma_\varepsilon^2}.$$

Letting $\lambda = {p\sigma_\varepsilon^2}/{\|\boldsymbol{\beta}\|^2}$, the value that minimizes the mean squared error function, the above equation states that the least squares estimator has a larger mean squared error than the ridge one when $2 + \lambda > 1$. The next section demonstrates logistic ridge regression for predicting violence.

6.3.6 Main Effects Ridge Logistic Regression Model

In Chapter 5, a main effects logistic regression model was found for the MacArthur Violence Risk Assessment Study (VRAS). Here, a logistic regression model is fit using the ridge estimator. The choice of λ is determined using cross-validation; based on the identified λ, this model is compared to the maximum likelihood model similar to that found in Chapter 5.

Chapter 1 introduced the VRAS and Chapter 5 discussed the data in detail; here `PCLtot` represents the overall PCL:SV scores, as opposed to the dichotomized `PCL` used in Chapter 5. A main effects logistic regression (MELR) and a ridge logistic regression (RLR) model are fit; for simplicity, these models are fit using only a subset of the continuous variables from VRAS dataset discussed in Chapter 5. Here, the variables are standardized to have variance 1 and mean-centered. All analyses are done in `R` and the code can be found in Appendix C. Table 6.3 displays the results of the MELR model.

Variable	$\hat{\beta}$	CI	
Intercept	-1.82	-2.04	-1.61
`PCLtot`	0.75	0.56	0.95
`BISnp`	-0.21	-0.41	-0.01
`BPRSa`	-0.27	-0.51	-0.05
`BPRSh`	0.36	0.11	0.61
`BPRSt`	-0.38	-0.66	-0.11
`PriorArr`	0.30	0.11	0.48
`SNMHP`	-0.21	-0.43	-0.01
`NASb`	0.33	0.13	0.52
`Function`	0.08	-0.12	0.28
`NegRel`	0.09	-0.10	0.26

Table 6.3: Estimated coefficients ($\hat{\beta}$) with confidence intervals (CI) for a main effects logistic regression model using standardized variables from the MacArthur Violence Risk Assessment Study (Monahan et al., 2001).

The logistic regression model is fit using maximum likelihood estimation; the negative

log-likelihood is the objective function to be minimized with respect to $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$:

$$\min_{\boldsymbol{\beta}} - \left(\sum_{i=1}^{N} y_i \sum_{j=0}^{p} x_{ij}\beta_j - \sum_{i=1}^{N} n_i \ln \left(1 + \exp \left\{ \sum_{j=0}^{p} x_{ij}\beta_j \right\} \right) \right),$$

where $x_{i0} = 1$ for all i. Similar to the linear regression objective function, a shrinkage penalty can be added to the logistic regression objective function:

$$\min_{\boldsymbol{\beta}} - \left(\sum_{i=1}^{N} y_i \sum_{j=0}^{p} x_{ij}\beta_j - \sum_{i=1}^{N} n_i \ln \left(1 + \exp \left\{ \sum_{j=0}^{p} x_{ij}\beta_j \right\} \right) \right) + \lambda \sum_{j=1}^{p} \beta_j^2.$$

Note that the shrinkage penalty does not include the intercept term, β_0.

A logistic regression model using ridge coefficients is examined; to determine the choice for the tuning parameter, λ, ten-fold cross-validation is conducted for values ranging from 0.0085 to 85. Recall that as $\lambda \to 0$ (or equivalently, as $\ln(\lambda) \to -\infty$), the coefficients approach the maximum likelihood estimates from Table 6.3; as $\lambda \to \infty$ ($\ln(\lambda) \to +\infty$), they approach 0. The value of the estimated coefficients across different values of λ are plotted in Figure 6.6; this is given on a logarithmic scale to better illustrate the trend. Starting from the ML estimates, the coefficients decrease toward zero as $\ln(\lambda)$ increases. The variable `BISnp`, the only MELR variable with a confidence interval containing zero, shrinks toward zero the fastest. All the other variables shrink at close to the same rate.

The optimal λ, in terms of minimized misclassification error using ten-fold cross-validation, was found to be $\lambda = 0.01$; this is displayed in Figure 6.7. Using this optimal λ, a ridge logistic regression model is constructed. The parameter estimates are given in Table 6.4 along with the ML estimates for comparison. Figure 6.8 plots both sets of estimates; Figure 6.9 plots the shrinkage of the estimates (the arrows represent the amount of shrinkage for the ML estimates to the ridge estimates).

The ten-fold cross-validated error for the ridge model is .179; for the main effects model, .185. This improvement results in six more correct predictions for the ridge logistic

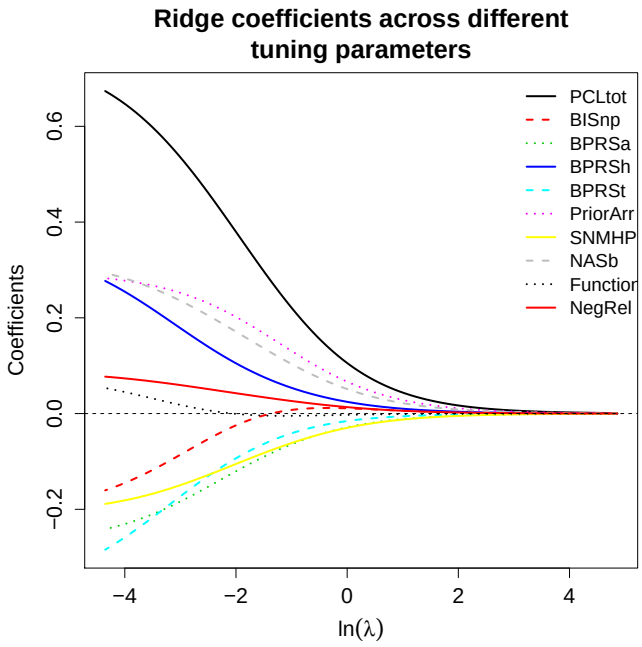

Figure 6.6: Ridge regression coefficients across different values of the natural logarthhm of the tuning parameter, $\ln(\lambda)$.

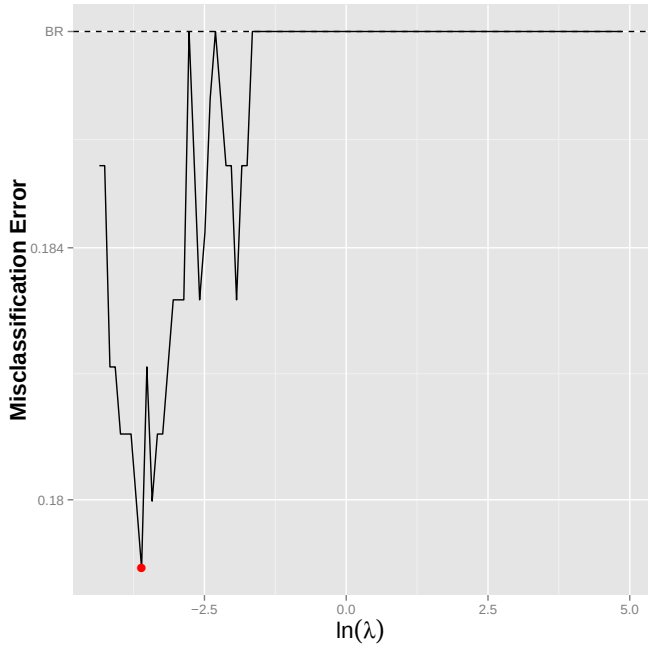

Figure 6.7: Ten-fold cross-validation for selecting the tuning parameter based on lowest misclassification error where the red dot (•) represents the minimum point.

Variable	$\hat{\beta}_{\text{ML}}$	$\hat{\beta}^R$
Intercept	-1.82	-1.71
PCLtot	0.75	0.61
BISnp	-0.21	-0.12
BPRSa	-0.27	-0.22
BPRSh	0.36	0.23
BPRSt	-0.38	-0.23
PriorArr	0.30	0.27
SNMHP	-0.21	-0.17
NASb	0.33	0.27
Function	0.08	0.03
NegRel	0.09	0.07

Table 6.4: Comparison of ridge estimators ($\lambda = .01$) and maximum likelihood estimators for logistic regression model using VRAS data (Monahan et al., 2001)

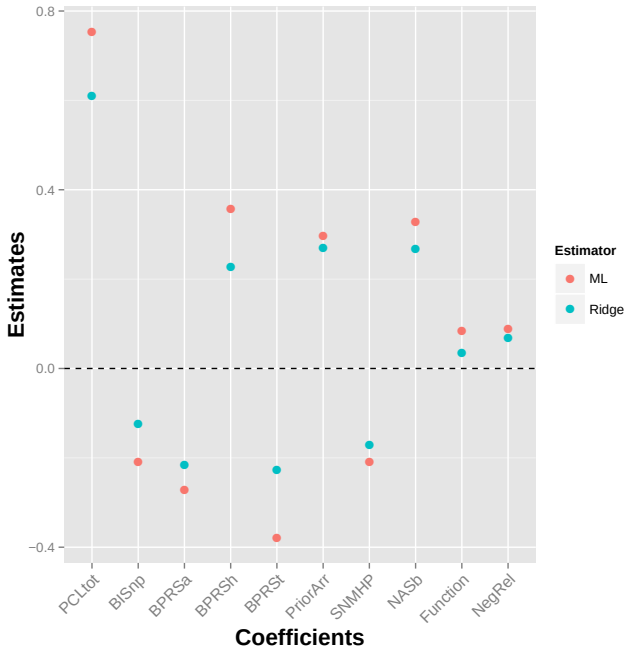

Figure 6.8: Comparison of ML and ridge logistic regression estimates for the VRAS data (Monahan et al., 2001).

regression model.

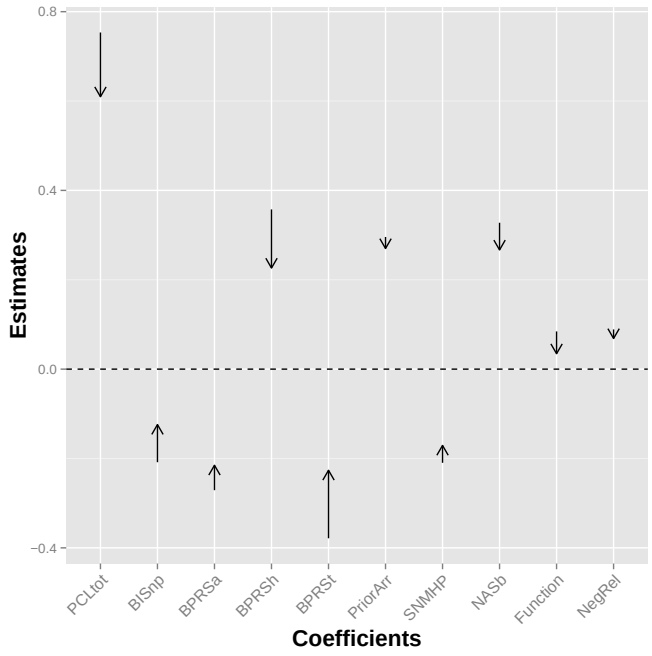

Figure 6.9: Amount of shrinkage for ridge logistic regression estimates for the VRAS data (Monahan et al., 2001).

6.4 Conclusion

The variance-bias trade-off is an important concept in prediction. The mean-squared error and its variants, such as the Brier Score, are ways to measure the overall error of a prediction method; these can be decomposed into the squared bias and variance of the estimator. Varying the complexity (or flexibility) of a prediction model generally leads to increased variance but decreased squared bias; the goal of increased accuracy in prediction is based on determining where the combination of the two are minimized (i.e., where the mean squared error is minimized).

Several types of shrinkage estimators designed to reduce the overall prediction error at the cost of increased bias were presented and demonstrated in this chapter. In every scenario presented, the biased estimator was shown to outperform the unbiased estimator in terms of mean squared error, especially when the number of observations is small. Even with a large number of observations, the superiority of the biased estimator is still present.

It should be clear that there is a great advantage to using biased estimators in prediction.

The property of bias in an estimator is one of the more well-known concepts in statistics; even in the most elementary courses it is usually discussed. Consequently, bias may be viewed as the most central criterion that an estimator should satisfy; but this is not necessarily the case. In prediction, an unbiased estimator should be discarded for one that leads to a more accurate prediction—without argument. Referring to Kelley's True Score estimator, Harold Gulliken, one of the influential figures in classical test theory, stated in his book *Theory of Mental Tests*, that "no practical advantage is gained from using [Kelley's] regression equation to estimate true scores" (p. 45). Hubert and Wainer (2012) bluntly reply to this remark, "[W]ho really cares about bias when a generally more accurate prediction strategy can be defined?" (p. 153).

Chapter 7

An Overview of Violence Prediction

"All models are wrong, but some are useful."

— George E. P. Box

This chapter looks at multiple studies using the Classification of Violence Risk (COVR) and Static-99R and Static-2002R actuarial tools. We conclude that within and across studies, the methods for predicting violent and dangerous behavior fail to satisfy the definition of clinical efficiency, cautioning their use in practice.

7.1 The Classification of Violence Risk (COVR)

The Classification of Violence Risk (COVR) introduced in Chapter 1 has been our main focus throughout. Recall that Monahan et al. (2001) used an iterative classification tree (ICT) scheme discussed in Chapters 1 and 5 to assign patients into five risk groups (see Table 7.1).

The ROC for the final ICT classification model provided an AUC of .88, which was noted to be significantly different from a chance accuracy of .50 (for a thorough discussion on the (mis)use of AUCs, see Chapter 4). Their model was not cross-validated (see Chapter 5 for a discussion of cross-validation). Table 7.2 displays a 2×2 contingency table for the results of the ICT model prediction of violence. Those predicted to be violent fell into the high and very-high risk categories.

Category	Risk	Point Estimate	95% CI
5	Very High	.762	[.654, .862]
4	High	.559	[.462, .653]
3	Average	.262	[.195, .324]
2	Low	.077	[.047, .111]
1	Very Low	.012	[.003, .024]

Table 7.1: The five risk categories for the Classification of Violence Risk (COVR) diagnostic test along with point estimate risks (in probabilities) and respective confidence intervals (CI) (Monahan et al., 2006).

		Violence		
		Yes (A)	No ($\bar{A}$)	Row Totals
Prediction	Yes (B)	105	60	165
	No ($\bar{B}$)	71	703	774
	Column Totals	176	763	939

Table 7.2: A 2×2 contingency table displaying the prediction of violence results using the ICT model presented in Monahan et al. (2001).

The base rate for predicted violence (i.e., the percentage predicted to be violent) is .175. In absolute terms, 165 patients were predicted by the ICT model to commit violence, 11 less than actually did (for a discussion of calibration, see Chapter 6). The sensitivity and specificity of this test is, respectively, .60 and .92; the positive and negative predictive values are, respectively, .64 and .91; the accuracy (i.e., the proportion correct) of the test is .86. The test meets the Bokhari-Hubert (BH) condition (see Chapter 3).

7.2 Validation Studies

Several studies have attempted to validate the COVR; this chapter takes a "meta-analytic" approach to assess just how well the COVR predicts violence. Although not as in depth, we also look at several studies examining the Violence Risk Appraisal Guide (VRAG) and the Static-99 (see Chapter 1 for a discussion of these measures). Our primary focus is on the COVR because it "exemplifie[s]" (Nadelhoffer et al., 2012, p. 68) actuarial methods for

predicting violence and is "user-friendly [and] produces excellent results" (Scurich & John, 2011, p. 83). The data used herein were (a) available directly from the article, (b) inferred from the article (i.e., raw numbers converted from percentages), or (c) requested from the authors.

7.2.1 Monahan et al. (2005)

In 2005, Monahan et al. conducted a study to validate the COVR actuarial tool they had developed earlier. Their data consisted of 157 patients from three hospitals in the United States who met the same inclusion criteria set forth in their original study (i.e., aged 18–40; English speaking; white, African-American, or Hispanic; diagnosed as having one or more disorders from a given list). Patients were classified as high risk (COVR score of 4 or 5), average risk (COVR score of 2 or 3), and low risk (COVR score of 1), and interviewed up to two times in a 20-week period after hospital discharge.

The 2×2 contingency Table 7.3 gives the validation study's results. A similar table was presented and discussed in Chapter 3 and Chapter 4; here, however, the data presented correspond to the authors' revised estimate of violence justified as follows:

> During qualitative review of the follow-up violence data, we realized that a number of the patients who had been classified as high risk by the software but who were not reported as violent during the follow-up (according to the strict operational definition given above) in fact presented strong evidence of violence. Indications that violence had actually taken place during the follow-up included violent acts that took place in an institution (for example, a jail or a hospital), evidence of violence several days after the 20-week follow-up window (as indicated by arrest records), and battery in which injury was highly likely but had been rated as "unknown." (p. 814)

The final follow-up sample consisted of 157 randomly selected patients from either the high-risk or low-risk groups, excluding the average-risk groups. The base rate for violence is

.23; the base rate for predicted violence is .35. The sensitivity of the test is .75; specificity is .77. The positive predictive value is .49 and the negative predictive value is .91. Although their revised estimate of violence improves their results (eight patients in the high-risk group previously classified as not violent were reclassified), the test still fails to meet the BH condition. In describing the results several years later, Monahan (2012) stated,

> The results of the COVR validation study were that during the follow-up approximately (a) one-in-ten of the patients classified by the instrument as "low" risk of violence committed a violent act; and (b) one-in-two of the patients classified by the instrument as "high" risk of violence committed a violent act. This difference was highly statistically significant. (pp. 192–193)

		State of Nature		
		A (Violence Present)	$\bar{A}$ (Violence Absent)	Row Totals
Prediction	B (Risk Present)	27	28	55
	$\bar{B}$ (Risk Absent)	9	93	102
	Column Totals	36	121	157

Table 7.3: A 2×2 contingency table for predicting violence risk among persons with mental disorders (Monahan et al., 2005).

7.2.2 Snowden et al. (2009)

Snowden, Gray, Taylor, and Fitzgerald (2009) "undertook the first independent study of the COVR" defined as a validation of the COVR not conducted by its original authors. The sample consisted of 52 forensic patients in two United Kingdom medium-security psychiatric units. The authors used three types of *inpatient* behavioral aggression as an outcome variable: verbal, property, and physical. Unfortunately, the authors do not provide definitions for these three types of aggression (and did not respond to an email inquiry about the definition of aggression used); the base rates of occurrence were, respectively, 1.00, .50, and .52. The COVR was significantly correlated ($p < .05$) with all three types of aggression.

The AUC measure for predicting property aggression was .57 and, for predicting physical violence, .73 (the latter value being significant, $p < .001$). These results led the authors to conclude that "the COVR proved to be a good predictor in this setting, and thus this study provides the first evidence of its usefulness in forensic services."

The base rate of violence, property and physical, is unusually high given other typical values in the literature. Focusing solely on property violence, it can be shown that, given COVR cutscores of 4, 3, 2, and 1, the COVR test fails the BH condition when using the two extreme cutscores (1 and 4) but not when using the middle cutscores (2 and 3). The accuracy for both these tests was .56. For physical violence, the COVR only fails to meet the BH condition when using a cutscore of 4. The highest accuracy was obtained with a cutscore of 2 (.69). When combining both property and physical violence, the base rate for violence increases to .65. None of the cutscores met the BH criterion; a cutscore of 1 or 2 produced the highest accuracy, .60.

Table 7.4 displays the results for the test predicting physical violence using 3 as a cutscore (i.e., patients with a COVR score of 4 or 5 are predicted to commit violence; patients with a COVR score of 1, 2, or 3 are not). The sensitivity of the test is .37; the specificity is .96. The positive predictive value is .91 and the negative predictive value is .59; both are larger than ½ implying satisfaction of the BH condition. The accuracy of the test is .65 and, as expected, is greater than the base rate for violence of .52.

		Physical Violence		
		Yes (A)	No ($\bar{A}$)	Totals
Prediction	Yes (B)	10	1	11
	No ($\bar{B}$)	17	24	41
	Totals	27	25	52

Table 7.4: Predicting physical violence using the COVR presented in Snowden et al. (2009).

7.2.3 Doyle et al. (2010)

Doyle, Shaw, Carter, and Dolan (2010) demonstrated the ineffectiveness of the COVR in predicting violence. Their data came from non-forensic acute mental health patients in the United Kingdom; it was the " first independent study of community violence to evaluate the validity of the COVR in a UK sample" (Doyle et al., 2010, p. 317). The authors followed a similar protocol for selecting their sample as in Monahan et al. (2001) and used the "slightly more inclusive measure of violence ... to control for any doubt regarding likely injury" (Doyle et al., 2010, p. 318), as was done in Monahan et al. (2005). The authors noted the nonsignificant AUC value of .58 and the nonsignificant X^2 of 7.89 with four degrees of freedom.

Their final sample included 93 patients, 22 of which committed at least one act of violence; thus, the base rate for violence in the sample was .24. Table 7.5 displays a 2×2 contingency table, with persons in the very-high-risk and high-risk groups predicted to commit an act of violence and persons in the other three are not. Only 5 patients were classified as high or very high risk of being violent and none committed any acts of violence; thus, the positive predictive value and the sensitivity of the test are 0. The specificity is .93; the negative predictive value is .75. The accuracy of the test is .71; the test fails to meet the BH condition.

		Violence		
		Yes (A)	No ($\bar{A}$)	Totals
Prediction	Yes (B)	0	5	5
	No ($\bar{B}$)	22	66	88
	Totals	22	71	93

Table 7.5: Predicting violence using the COVR presented in Doyle et al. (2010).

7.2.4 McDermott et al. (2011)

McDermott, Dualan, and Scott (2011) looked at the predictive ability for the COVR

diagnostic test among civilly committed psychiatric patients at a United States forensic facility. The study used 146 patients committed to the facility as not guilty by reason of insanity or mentally disordered offender. The authors' outcome variable is physical aggression defined as "any physical contact initiated by a patient wherein the intent was to do physical harm, such as pushing, kicking, or biting or using a weapon to threaten others." (McDermott et al., 2011, p. 431). McDermott et al. (2011) further noted that "Unlike the MacArthur Violence Risk Assessment Study, in our study acts of physical aggression did not necessarily result in physical injury" (p. 431). Each patient was part of the study for twenty weeks following the completion of the COVR. The authors concluded the COVR was an efficient diagnostic tool for aggression prediction based on a significant area under the curve (AUC) measure and a chi-squared analysis indicating that "COVR scores were strongly associated with overall aggression" (McDermott et al., 2011, p. 431).

Assuming high and very high COVR scores are indicators for aggression, Table 7.6 gives the 2×2 contingency table for the results of COVR and aggression prediction. The base rate for aggression in the sample is .15; the accuracy of the COVR test for the sample is .84, failing to outperform base-rate prediction (of .85). For this test, the sensitivity and specificity are, respectively, .41 and .91; the AUC is .66. The PPV and NPV are, respectively, .45 and .90.

		Aggression		
		Yes (A)	No ($\bar{A}$)	Totals
Prediction	Yes (B)	9	11	20
	No ($\bar{B}$)	13	113	126
	Totals	22	124	146

Table 7.6: Predicting aggression using the COVR presented in McDermott et al. (2011).

7.2.5 Sturup et al. (2011)

In a another study, 331 patients from psychiatric hospitals in Sweden were interviewed before they were discharged and violent behavior was measured up to twenty weeks after their release (Sturup et al., 2011). Nineteen patients displayed violent behavior (four were recorded in the criminal register as violent crimes); the base rate for violence in this sample is .06. Violence was defined as given in Monahan et al. (2001) and reported violence from criminal activity from the criminal register were defined as "aggravated assault, assault, violence or threat to a public servant and violently resisting arrest (there were no homicide, rape or other felonies recorded during the follow-up period for anyone in the sample)" (Sturup et al., 2011, p. 162).

The COVR software classified 191 patients as very low risk, 92 as low risk, 37 as average risk, 7 as high risk, and 4 as very high risk. Table 7.7 summarizes the classifications with the number of patients displaying violent behavior. The authors reported a significant AUC of .77; the ROC curve is shown in Figure 7.1. Regardless of which risk level is used as a cutscore for predicting violence, the test fails to meet the BH condition.

		Violent Behavior		
		Yes (A)	No ($\bar{A}$)	Totals
Risk	Very High	2	2	4
	High	2	5	7
	Average	7	30	37
	Low	4	88	92
	Very Low	4	187	191
	Totals	19	312	331

Table 7.7: Predicting violent behavior using the COVR presented in Sturup et al. (2011).

Table 7.8 displays the positive and negative predictive values, their average, the accuracy, and the AUC for a diagnostic test using each cutscore. The largest AUC is for a cutscore of 2 (this is also the "best" cutscore according to the distance measure D and Youden's J [see Chapter 4]), suggesting a COVR score in the average-to-very-high risk categories is a

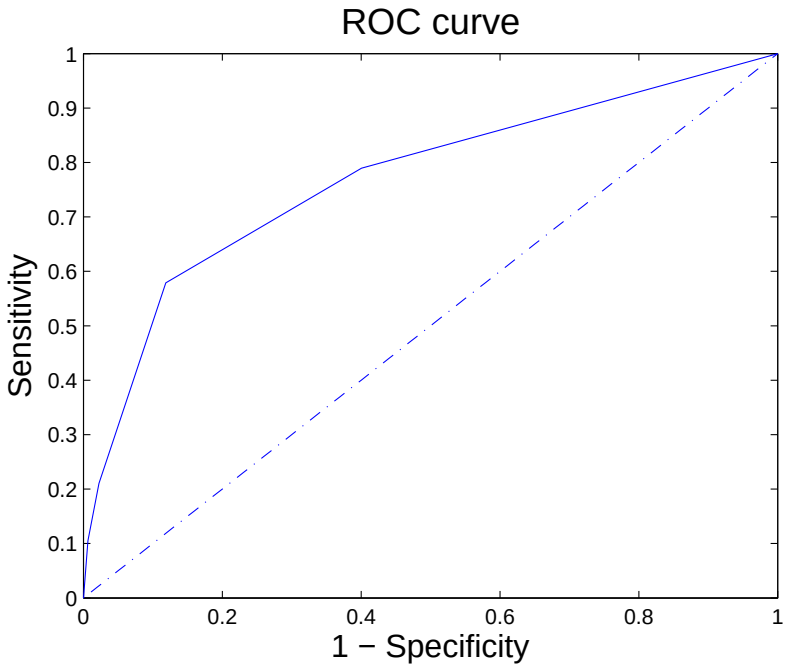

Figure 7.1: ROC curve for COVR validation study (Sturup et al., 2011).

good indicator for someone who will commit violence. Using this cutscore, the accuracy of the test is .86 and the PPV is .23 (37 out of 48 patients would be incorrectly predicted to be at risk for committing violence). The PPV and NPV, taken together, suggest a cutscore of 4 is the best, implying that anyone in the very-high-risk category is at considerable risk. This test provides an accuracy exactly equal to the base-rate probability $P(\bar{A})$. The AUC for this test, however, is the lowest of all four cutscores; the cutscore is also the worst in terms of the distance measure D and Youden's J.

Cutscore	PPV	NPV	Accuracy	AUC
4 (High)	0.50	0.95	0.94	0.55
3 (Average)	0.36	0.95	0.93	0.60
2 (Low)	0.23	0.97	0.86	0.73
1 (Very Low)	0.11	0.98	0.61	0.69

Table 7.8: Positive predictive value (PPV), negative predictive value (NPV), accuracy, and AUC for four cutscores for the COVR.

7.3 Aggregation of Studies

There are three obvious problems with combining the five validation studies. The first is that the definition of violence was unclear in one study (Snowden et al., 2009) and slightly different in another (Sturup et al., 2011). The second is that one study (Monahan et al., 2005) only included high risk and low risk patients, excluding all patients scoring a 2 or 3 on the COVR. The third is that the aggregated population is heterogeneous: three studies included recently discharged patients (Monahan et al., 2005; McDermott et al., 2011; and Doyle et al., 2010) whereas two studies included hospitalized patients (Snowden et al., 2009; and Sturup et al., 2011); two studies included patients from the United States (Monahan et al., 2005; and McDermott et al., 2011), two from the United Kingdom (Snowden et al., 2009; and Doyle et al., 2010), and one from Sweden (Sturup et al., 2011). Despite these caveats, the five studies were combined.

Table 7.9 displays a 2×2 contingency table for predicting violence in the five studies. In this table, patients scoring a 4 or 5 on the COVR were predicted to commit an act of violence (or aggression) and those scoring a 1, 2, or 3 were not. The base rate for violence is .16; the base rate for violence prediction is .13. The accuracy of the aggregate sample is .83, below the mark (i.e., less than $P(\bar{A}) = .84$). The sensitivity and specificity are, respectively, .39 and .92; the AUC is .66. The PPV and NPV are, respectively, .48 and .89. Aggregated over the five validation studies, the COVR fails to meet the BH condition.

		Violence		
		Yes (A)	No ($\bar{A}$)	Totals
Prediction	Yes (B)	50	54	104
	No ($\bar{B}$)	76	599	675
	Totals	126	653	779

Table 7.9: Predicting violence using the COVR across all five validation studies.

Table 7.10 displays a 2×2 contingency table, but this time excluding those who scored a 2 or a 3, as done in Monahan et al. (2005). The data were not directly available for

McDermott et al. (2011). From their paper it could only be determined that "six of the 84 (7%) patients who scored in the low or very low categories committed an act of aggression" (McDermott et al., 2011, p. 432). Because this description includes patients scoring a 2 on the COVR, it was indeterminate just how many scored a 1 (an email to the corresponding author did not receive a response); thus, none of the nineteen patients scoring a 1 on the COVR were predicted to be violent. The base rate for violence is about the same as the all-inclusive aggregation (.16); however, the base rate for violence prediction is greatly increased (.22). The sensitivity is improved, but the specificity worsens (.65 and .86, respectively); the AUC shows a large increase (.75). The PPV is the same; the NPV is improved (.92). The accuracy is slightly worse (.83) and the COVR still fails to meet the BH condition.

		Violence		
		Yes (A)	No ($\bar{A}$)	Totals
Prediction	Yes (B)	50	54	104
	No ($\bar{B}$)	27	339	366
	Totals	77	393	470

Table 7.10: Predicting violence using the COVR across all five validation studies (excluding patient scoring a 2 or 3).

Table 7.11 displays a summary of the five studies. Because the sensitivity and specificity are independent of base rates, we can calculate the expected positive and negative predictive values; these are included in the table (denoted $\mathbb{E}$(PPV) and $\mathbb{E}$(NPV), respectively).

	Monahan et al. (2005)	Snowden et al. (2009)	Doyle et al. (2010)	McDermott et al. (2011)	Sturup et al. (2011)
N	157	52	93	146	331
Violence Base Rate	.23	.52	.24	.15	.06
Violence Selection Ratio	.35	.21	.05	.14	.03
Reported AUC	.70	.73	.58	.73	.77
2×2 AUC	.76	.67	.46	.66	.59
Sensitivity	.75	.37	.00	.41	.21
Specificity	.77	.96	.93	.91	.98
PPV	.49	.91	.00	.45	.36
$\mathbb{E}$(PPV)	.67	.88	.68	.55	.30
NPV	.91	.59	.75	.90	.95
$\mathbb{E}$(NPV)	.89	.70	.89	.93	98
Accuracy	.76	.65	.71	.84	.93
O_B	0.96	10.00	0.00	0.82	0.57
$O_{\bar{B}}$	0.10	0.71	0.33	0.12	0.05
OR	9.96	14.12	0.00	7.11	11.62
RR	5.56	2.19	0.00	4.36	7.76
DLR_B	3.24	9.26	0.00	4.61	9.38
$DLR_{\bar{B}}$	0.33	0.66	1.08	0.65	0.81
Bokhari-Hubert Condition Met?	No	Yes	No	No	No

Table 7.11: Summary of the five COVR-validation studies.
N is the sample size; AUC is area under the ROC curve; PPV is positive predictive value; NPV is negative predictive value; O is odds; OR is odds ratio; RR is relative risk; DLR is diagnostic likelihood ratio.
Note: aside from *Reported AUC*, all statistics are calculated using data in the form of 2×2 contingency tables where a COVR score of 4 or 5 leads to a prediction of violence and a score of 1, 2, or 3 does not.

7.4 Sexual Recidivism

7.4.1 Static-99R & Static-2002R

Chapter 1 introduced the Static-99 and Static-2002 in detail. Both these measures were revised due to the possibility of overestimating recidivism for older males (Helmus, Thornton, et al., 2012); the revised estimates are the Static-99R and Static-2002R, respectively. For both scales, the age range was categorized by four groups: Aged 18–34.9; aged 35–39.9; aged 40–59.9; aged 60 or older. For the Static-99R, the scoring for the four groups is, respectively 1, 0, −1, and −3; for the Static-2002R the scoring is, respectively, 2, 1, 0, −2. The low group for the Static-99R are scores ranging −3 through 1; nothing else is changed. For the Static-2002R the scoring is as follows: High (9 or higher); moderate-high (7 or 8); moderate (5 or 6); low-moderate (3 or 4); and low (−2 through 2). Table 7.12 displays the Static-99R and Static-2002R scores and relative risk/hazard ratios associated with each score. On the Static-99's website, an interpretation guide is provided for these scores. As an example, for a score of 6 the following interpretation is provided: "The recidivism rate of sex offenders with the same score as Mr. XXX would be expected to be approximately 2.6 times higher than the recidivism rate of the typical sexual offender (defined as a median score of 3)" (see *Static-2002R relative risk ratio table*, Static-99, 2013).

The new norms established for the revised Static-99 and Static-2002 provides a range of recidivism risk, rather than a single percentile. According to Helmus et al. (2009)

> Currently, our recommendation is to report recidivism estimates with the new norms in two stages. The first stage involves reporting an empirically-derived range of recidivism risk. The recidivism estimates from the [Correctional Service of Canada] samples represent the lower bound of the range and the preselected high-risk samples are the upper bound of the range. ... The second stage involves making a professional judgment as to where a particular offender is likely to fall within that range. This judgment represents a separate task from reporting the empirical re-

Static-99R		Static-2002R	
	Relative Risk		Hazard
Score	Ratio	Score	Ratio
10+	8.47		
9	6.48	9+	6.90
8	4.96	8	5.00
7	3.80	7	3.62
6	2.91	6	2.63
5	2.23	5	1.90
4	1.71	4	1.38
3	1.31	3	1.00
2	1.00	2	0.72
1	0.77	1	0.52
0	0.59	0	0.38
−1	0.45	−1	0.28
−2	0.34	−2	0.20
−3	0.26		

Table 7.12: Scoring distribution and their associated hazard ratios for the Static-99R and Static-2002R.

> cidivism rates; currently, there is no research to assess how well evaluators are able to make this judgment. Until further research is conducted, however, this professional judgment is unavoidable. It is also important to note that regardless of the evaluator's opinion of which sample the offender most closely resembles, recidivism rates of both samples should be reported in all cases. Although reporting absolute recidivism rates as a range may appear less precise, it is likely more realistic given that predicting behavior was likely never as simple as associating a single number with a single Static-99 score.

Although we agree with the final sentence (i.e., a range of absolute recidivism rate is more realistic than a single number) we do not agree that the second stage is necessary or even appropriate. It is certainly avoidable, and until further research is conducted, no judgment should be made as to where an individual falls within the estimated range.

7.4.2 Validation Studies

In developing the revised edition of the Static-99R and Static-2002R, Helmus, Thornton, et al. (2012) combined 24 studies (see referenced article for citations of these studies); all 24 studies included Static-99 data ($N = 8,390$; 23 recorded sexual recidivism and 19 recorded violent recidivism) and seven also included Static-2002 data ($N = 2,609$). Eleven samples were from the Canada, six from the United States, two from the United Kingdom, one from Denmark, one from Austria, one from Germany, and one from New Zealand. The sexual offenders across these studies ranged between the ages 18 and 84 and were released between the years 1957 and 2007. We note a report using 20 of the 24 studies was published (Babchishin, Hanson, & Helmus, 2011).

Table 7.13 displays results of combined 24 studies. This table is reconstructed from the table provided in the appendix (Helmus, Thornton, et al., 2012, p. 95). Note the values provided in the original article are given in terms of percentages; the values provided in Table 7.13 are raw numbers. Although these raw numbers were estimated as best as possible, there appear to be some discrepancies. For example, the authors state the 5-year sexual recidivism rate for patients scoring low on the Static-2002 is 3.7%; the total number of patients scoring low on the Static-2002 is given as 686. Simply multiplying 686 by .037 gives 25.382, or simply 25 after rounding. However, 25 is only 3.6% of 686 (when rounded accordingly to one decimal) and 26 is 3.8%; neither matching the authors' number. This was the case for several entries so some estimates may be off slightly, but not by more than one or two per cell, a meaningless difference in terms of the overall conclusion. In addition, because of the large sample sizes within each Static-99 level, the raw numbers to match the correct percentages were not unique. For instance, 4.1% of the 2,380 patients scoring low on the Static-99 could translate to 97 or 98 patients. Again, the estimates may only be off by one or two per cell.

The area under the curve for the Static-99R is .71 and is statistically significant from

Static-99

Score	Sexual Recidivism: Yes	Sexual Recidivism: No
High	308	903
Moderate-high	265	1565
Moderate-low	184	2501
Low	97	2283
Column Totals	854	7252

Static-99R

Score	Sexual Recidivism: Yes	Sexual Recidivism: No
High	326	898
Moderate-high	248	1589
Moderate-low	166	2043
Low	114	2722
Column Totals	854	7252

Static-2002

Score	Sexual Recidivism: Yes	Sexual Recidivism: No
High	72	155
Moderate-high	112	332
Moderate	94	620
Moderate-low	44	649
Low	19	512
Column Totals	341	2268

Static-2002R

Score	Sexual Recidivism: Yes	Sexual Recidivism: No
High	64	122
Moderate-high	98	271
Moderate	94	556
Moderate-low	60	658
Low	25	661
Column Totals	341	2268

Table 7.13: Results from 24 combined studies using the Static-99 and Static-2002 (with revised versions included) for predicted sexual recidivism after five years.

chance prediction. The ROC plot for these data are shown in Figure 7.2. Suppose we recommend that all patients scoring high on the Static-99R remain locked up. Our results can then be put in a 2×2 contingency table, as shown in Table 7.14. The AUC for this test is .63; the odds ratio is 4.39; the relative risk is 3.48; the diagnostic likelihood ratio for positive predictions is 3.09; the diagnostic likelihood ratio for negative predictions is .70; the accuracy is .82; a Pearson's chi-square test results in $X^2 = 396.4$, $p < .0001$. All these point to a very good test. The base rate for sexual recidivism (after 5 years) is .11; if we recommend no patients to be locked up, our accuracy increases to .89. Instead, using our "very good test" we mistakenly lock up 898 of the 1224 patients we recommended remain incarcerated. The positive predictive value for this test is .27; the test fails to meet the BH

condition, in fact, it is not even close. Using the Static-2002R data, the results are not much better (PPV = .34).

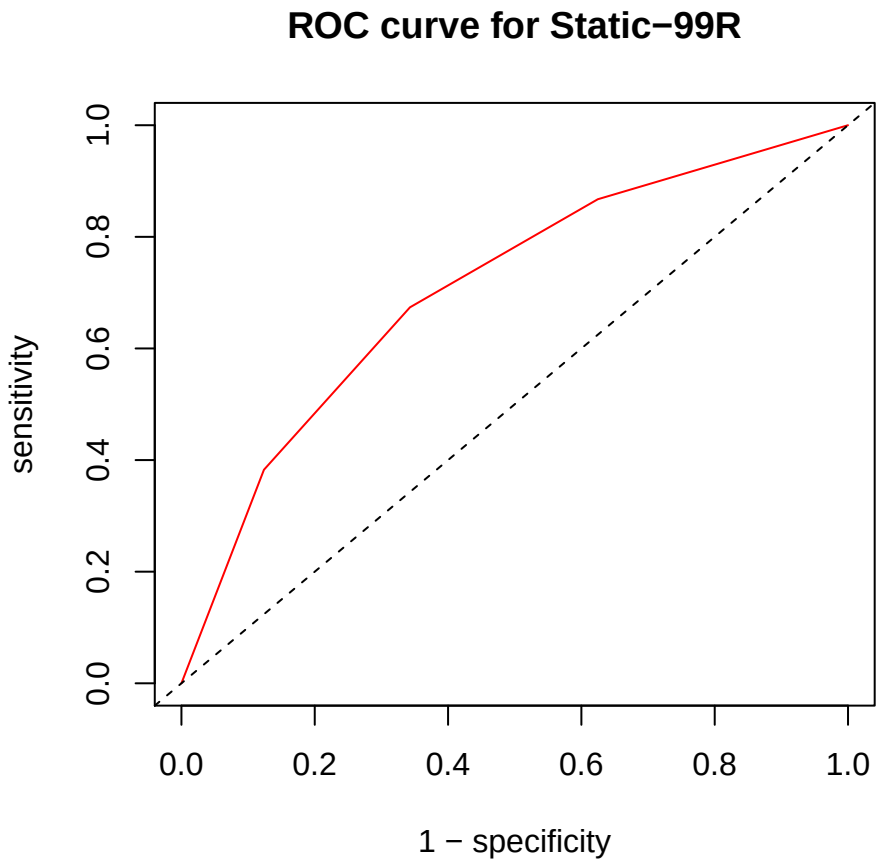

Figure 7.2: ROC curve for Static-99R (see Table 7.13)

		Sexual Recidivism	
		Yes	No
Predicted	Yes	326	898
Recidivism	No	526	6356

Table 7.14: A 2×2 contingency table for predicting sexual recidivism using the Static-99R. A score of 6 or higher leads to a prediction of recidivism.

As previously mentioned, Helmus et al. (2009) recommends that a judgment be made once a range is determined by the Static-99R, as to where the individual falls within this

range. In *State of New Hampshire v. William Ploof* (2009), the court ruled this clinical judgment does not meet the requirement of the RSA 516:29-a I(b) (N.H. Stats., 2014b, see Chapter 1); however, the court eventually concluded that the Static-99R is admissible in the court of law. These conclusions appeared to be largely influenced by the accuracy of the Static-99R as deemed by the AUC values. There was no mention of positive predictive values from either side. Had the Court been told that only about one in four individuals scoring high on the Static-99R actually recidivate after five years, perhaps their conclusion would be different. The Court also noted that "evidence like [the Static-99R] is particularly helpful to a jury in this sort of case because it helps the jury differentiate between sexual offenders and determine whether a particular respondent is 'likely to engage in acts of sexual violence if not confined in a secure facility'" (*State of New Hampshire v. William Ploof*, 2009, p. 26). With all due respect to the Court and its decision, we disagree; although an individual who scores in the high range on the Static-99R may be more likely to recidivate than one who does not, a PPV of .27 is not an indication that the individual is "likely" to recidivate.

7.5 Conclusion

The results in Table 7.11 show a number of things. First, the measures of predictive accuracy (sensitivity, specificity, AUC, positive and negative predictive values, and accuracy) are all generally lower in the validation studies (there are some exceptions; e.g., the reported PPV is higher in Snowden et al., 2009). The positive predictive value is generally much lower than expected, given the base rate of the sample and the original study's reported sensitivity and specificity; the negative predictive value is more close to its expected value. Depending on the statistics used, any one of the five tests can be made to look impressive (except for Doyle et al., 2010) but the fact that only one validation study outperforms base-rate prediction is telling: The COVR simply fails to predict violence with acceptable accuracy.

Others flaws in several of these studies have also been pointed out (e.g., see Large, Ryan, & Nielssen, 2010; McCusker, 2007)

The original data used to construct the COVR provided very good results and outperformed base rate prediction but as the results of the five validation studies suggest, this may have been due to overfitting (evidence for overfitting is also provided in Chapter 5).

Chapter 8

Conclusion

"Statisticians, like artists, have the bad habit of falling in love with their models."

— George E. P. Box

This thesis is concerned with prediction in the social sciences. Specifically, it focuses on four quantitative aspects of prediction, primarily in the realm of predicting dangerous and violent behavior.

The first topic area involved the notion of clinical efficiency as it relates to a prediction tool or diagnostic measure; this was presented in Chapter 3. Clinical efficiency as defined by Meehl and Rosen (1955), is when a diagnostic instrument outperforms simple base rate prediction. A general condition for determining when clinical efficiency is met was presented, called the Bokhari-Hubert (BH) condition; it has several possible reinterpretations in terms of odds ratios, relative risks, diagnostic likelihood ratios, and the Goodman-Kruskal lambda coefficient. As shown, many diagnostic tools in the area of violence prediction fail to meet this criterion; a major implication of this failure is that when the base rate is less than one-half, a positive prediction is more likely to be incorrect.

A second topic of interest developed in Chapter 4, concerned the use of the receiver operating characteristic (ROC) curve and the related measure of accuracy, the area under the ROC curve (AUC). The AUC is a commonly used measure of accuracy for evaluating prediction tools, such as those designed to predict violent behavior, and is often touted for being independent of base rates. To the contrary, however, it is our contention that base rates should be taken into consideration in evaluating the actual usage of a test, particularly when the base rate is close to zero; not doing so (i.e., relying solely on the AUC) is misleading.

When presenting the results of a prediction tool, alternative measures such as the positive predictive value that take into account the base rate should be included.

The third area discussed was cross-validation. Chapter 5 uses several state-of-the-art statistical learning methods for predicting violence. Although the methods are powerful, like any predictive model they are susceptible to overfitting. Cross-validation is a crucial aspect in the construction of any prediction model; this chapter demonstrates that inflated and spurious results often arise when a model is not cross-validated. One specific result emphasized is that when base rates are low, positive predictions are infrequent unless overfitting occurs or the cost ratio of false negatives to false positives is increased; the latter is ethically questionable when predictions of dangerous behavior assist in decisions such as preventive detention or the imposition of the death penalty.

A fourth topic deals with what is commonly referred to as the bias-variance trade-off. Generally, as models become more complex, the squared bias decreases as the error variance increases. Chapter 6 examines this phenomenon and advocates for the use of biased estimators over those that are unbiased when the mean squared error is smaller. For example, a ridge regression estimator will always have less error variance compared to least squares when used on new data, despite the least squares estimator being unbiased; a consequence of this result is that the mean squared error for the ridge estimator may be less than for least squares. In addition, the relationship between the James-Stein estimator and Kelley's True Score estimation is shown. Several illustrative examples demonstrate the improvement in mean squared error using biased estimators over their unbiased counterparts.

8.1 Prediction in Other Areas

This thesis has focused primarily on predicting violent and dangerous behavior, but the topics developed extend to all areas concerned with prediction. One very important field of application is medicine, and specifically with screening for disease and cancer. Over-

diagnosis in medicine is a heavily debated topic (e.g., see Welch, Schwartz, & Woloshin, 2011); one reason that the medical community may be overdiagnosing patients is because of wide-spread programs for screening healthy individuals (i.e., screening a population with a probable low base rate for the disease of interest). There are many areas where unnecessary screening may be present (Cassels, 2012; Welch et al., 2011), but probably the two most notorious involve prostate cancer and the prostate specific antigen (PSA) exam and breast cancer screening with mammographies. The latter recent received attention when A. B. Miller et al. (2014) made national headlines with their publication of a twenty-five year follow-up study comparing breast cancer incidence and mortality among nearly 90,000 Canadian women aged 40–59 who did or did not have mammography screenings. No reduction in mortality was found for those screened; rather, screening led to 142 more false positives of breast cancer in the women screened.

Overdiagnosis leads to unnecessary and costly treatment and stress; the epitome of this is the full-body scan (Cassels, Van Wiltenburg, & Armstrong, 2009). Furtado et al. (2005) report that in a sample of 1192 patients (903, or 76%, of whom were self-referred), 1,030 (86%) had at least one abnormal finding discovered by computed tomographic (CT) screening, with an average of 2.8 abnormalities per patient. In addition, 445 patients (37%) were recommended for further evaluation; a tumor found when one is not looking has been called an *incidentaloma*. According to some, most incidentalomas are not cancerous (Welch et al., 2011).

Another area of screening gaining popularity is positron emission tomography (PET) scans for diagnosing Alzheimer's disease. One might question what happens when an incorrect diagnosis of Alzheimer's is made. Analogous to full-body scans are whole genome scans (National Public Radio, 2013b); some researchers suggest adopting genetic information with other factors to predict future drug use in adolescents (Conner, Hellemann, Ritchie, & Noble, 2010).

Prenatal screening is another controversial idea in screening. NPR reported on a new

blood test for detecting Down syndrome in a fetus (R. Stein, 2014). Two statements made in the article should concern any reader versed in the topics presented in this thesis. The first is, "Doctors recommend that all pregnant women get screened for Down syndrome and other trisomies" (R. Stein, 2014, para. 5). The second appears several paragraphs later: "The test's 'positive predictive value,' which is its ability to accurately predict whether the fetus has Down syndrome, was 10 times greater than standard testing, the researchers reported (45.5 percent compared with 4.2 percent)" (R. Stein, 2014, para. 10). Although it is commendable that the positive predictive value was reported by a popular media outlet, the article fails to emphasize what this implies. If all pregnant women are screened for Down syndrome in their unborn child and over half of the predictions of Down syndrome are incorrect, what might this suggest about the aborting of healthy babies?

The issue of overdiagnosis seems related to innumeracy, discussed in Chapter 2 along with the two informative papers by Gigerenzer et al. (2007) and Monahan (2007). Overdiagnosis is not limited to physical diseases; Zimmerman et al. (2009) report that in a study of 480 psychiatric outpatients completing the Mood Disorders Questionnaire (designed to diagnose bipolar disorder), the positive predictive value was .34, suggesting that almost two-thirds of those diagnosed as being bipolar were not. Other examples gaining a lot of attention in the media are the excessive diagnosis of attention deficit hyperactivity disorder (ADHD) and a new "condition" called sluggish cognitive tempo that includes children failing to the meet the ADHD criterion (A. Schwarz, 2014).

Researchers at the Cincinnati Children's Hospital are attempting to identify suicide risk based on language used in notes of those who have committed suicide (National Public Radio, 2013a). According to the head of research team, Dr. John Pestian, when suicidal children come to their emergency room, it would be of value to know which kids should receive help externally and which should receive help from within the hospital; the claim is that their method is about 90% accurate in predicting who is suicidal (National Public Radio, 2013a). The important number to note here is the probability of a false negative.

Substantial discussion is available about brain trauma caused by head injuries among football athletes in the National Football League (NFL). The problem is apparently so bad that a lawsuit has been filed on behalf of numerous former players. The diagnosis in question is chronic traumatic encephalopathy (CTE), although currently there is no way to accurately diagnose it. The urgency of the matter has prompted a large investment of resources into CTE diagnosis, and in turn, to various diagnostic tests not as valid as claimed (Belson, 2013).

Also in sports, Major League Baseball (MLB) has undergone a serious blow to its image caused by public awareness of players using performance enhancing drugs (PEDs). This has led to increased punishments for perpetrators; of concern for MLB is the diagnostic test's sensitivity—MLB wishes to find as many (preferably all) PED users as possible. For players (and its powerful union), the positive predictive value is of major concern—when a player tests positive, how likely is it that he actually used PEDs? Because MLB and the players' union must reach agreements about testing procedures, these clearly are issues relevant to the discussion. The problems associated with drug testing also exists in other sports, most notably in cycling and track and field. The International Olympic Committee (IOC), for example, allows an "expert panel" to declare female athletes ineligible for competition when their levels of testosterone are "within the male range" (International Olympic Committe, 2014); however, Healy, Gibney, Pentecost, Wheeler, and Sonksen (in press) found that although mean testosterone levels may differ, there is much overlap between male and female levels in elite athletes and suggest that the IOC's decision to "limit participation in elite events to women with a 'normal' serum testosterone is unsustainable."

Similar to testing athletes for PEDs, alcohol and drug screening are subject to the same issues described in this thesis. Urine drug screening is the common way to detect drug use; although these types of tests perform well for some drugs, such as marijuana and cocaine, they are fallible and perform markedly worse for other drugs (Moeller, Lee, & Kissack, 2008). Koerth-Baker (2014) states that field sobriety tests for determining when

the driver of a vehicle is under the influence of alcohol "has been shown to catch 88 percent of drivers under the influence of alcohol" (para. 2), but what does this mean? Clearly it does not mean that field sobriety tests catch 88 percent of all drivers who are drunk. With the legalization of marijuana in two states (Washington and Colorado), Koerth-Baker (2014) discusses law enforcement's (in-)ability to detect drivers under the influence of marijuana.

Given the recent research misconduct in psychology, groups of researchers are beginning to specialize in detecting fraudulent studies. Social scientists generally agree that this is important, but incorrectly accusing someone of fraud could have devastating effects for one's career and thus, the properties of such detection need close consideration. Audits by the IRS and other businesses is an area where prediction should be carefully evaluated, as is prediction of political elections or the potential impacts of climate change. Or, consider credit risk calculators used by banks and other loan agencies. Areas of prediction where the ideas presented in this thesis are relevant are wide-spread, indeed.

Consider the long-sought method for accurately detecting lies. Lie detectors are vulnerable to many errors, making accurate lie detection an extremely difficult task Vrij, Granhag, and Porter (2010). In a meta-analysis examining people's ability to discriminate lies from truth, Bond and DePaulo (2006) estimated that the average accuracy, across more than two-hundred studies, was about 54%, slightly better than chance. They also noted that lies were correctly classified only 47% of the time (61% of truths were correctly identified). The United States Transportation Security Administration (TSA) reportedly spent around one billion dollars training TSA officers to identify terrorists based on nonverbal cues, such as facial expressions (Tierney, 2014). According to the *New York Times* report, only one percent of the more than 30,000 passengers suspected each year are arrested, and the arrests were not linked to terrorist plots. Based on these results and those from the Bond and DePaulo study, the TSA recommended cutting funds from the program (Tierney, 2014).

Eyewitness recall, like lie detection, is notoriously inaccurate. To evaluate the accuracy of eyewitness recall, Wixted and Mickes (2012) suggest the use of receiver operating

characteristic curves. These authors state that the probative value—defined as the sensitivity divided by the false positive rate that was given as the diagnostic likelihood in Chapter 3 (DLR_B)—is "an *irrelevant* [original emphasis] consideration when trying to decide which of two lineup procedures is better" (p. 276). The authors' arguments are not necessarily invalid but their support for the ROC curve suffers from the same criticisms presented in Chapter 4.

Researchers at Google introduced a "big-data" method for predicting influenza outbreaks based solely on search queries entered into Google's search website (Ginsberg et al., 2009)—they called this *Google Flu Trends* (GFT). When tested on new data, the model showed a mean correlation of .97 with the United States' Centers for Disease Control and Prevention (CDC) results, based on "influenza-like illness" visits to the physician (Ginsberg et al., 2009). Only a couple years later it was reported that GFT was consistently overestimating the number of flu cases. As big data becomes ingrained into everyday vocabulary, it is important to remember that big-data methods are susceptible to many of the issues discussed in this thesis. The GFT anecdote is also a reminder that researchers should continually be updating, validating, and improving a prediction model.

Another example of model overestimation comes from medicine and the diagnosis of high cholesterol (Kolata, 2013). An online calculator intended to provide interested patients with their risk of heart attack or stroke, was found to consistently overestimate the risk; the diagnosis of high risk often leads to prescribing cholesterol-lowering statins. The data are based on the Framingham Heart Study; the extrapolation of the 1980's Framingham population to today appears faulty.

When predictions go wrong, should someone be held accountable? The Italian judicial system certainly thought so when seven Italian seismologists were convicted of manslaughter and sentenced to prison for *not* predicting a deadly earthquake (Povoledo & Fountain, 2012). This is a rather dramatic case, but certainly some accountability should be taken when predictions (or lack thereof) have significant impact on numerous lives. When responsibility

is assumed, it is hoped that careless and unethical prediction may be less common.

8.2 Concluding Remarks

To say that violence is a serious problem in the United States is quite an understatement; as a society we incur far too many mass shootings. One of the first questions people ask without fail, is how could this have been prevented (i.e., why was this not predicted), and what was the shooter's mental health status. After the April 2014 Fort Hood mass shooting, the second such incident at Fort Hood in five years, the topic of Ivan Lopez's mental health was immediately brought to attention (Shapiro, 2014); this is the norm (e.g., see Memmott, 2014).

Media outlets report that mass shootings are increasing (e.g., see Follman, 2012b; George, 2013; Plumer, 2012b; Strasser, 2014); even Attorney General Eric Holder suggests this (The Associated Press, 2013). This conclusion, however, may need to be more nuanced depending on how one defines a mass shooting or examines the data (Fox, 2012; Plumer, 2012a; J. Walker, 2014). Mass shootings are a terrible thing, but the media's pandering to people's fear is reminiscent of the "superpredator" debacle in the 1990s. Dilulio (1995) used the term superpredator to describe young adolescents raised under certain conditions who are "perfectly capable of committing the most heinous acts of physical violence for the most trivial reasons" (p. 23). The superpredator scare garnered much media attention, driven by Dilulio's and others' predictions that there would be dramatic increases in crimes committed by these superpredators (Howell, 2009). Their predictions never amounted to anything except generating a lot of unnecessary fear and panic (*The New York Times*, 2014).

In the presence of degrading accuracy, Berk (2012) suggests that revising one's prediction tool is necessary. One issue with this idea and specifically with respect to predictions of violent or dangerous behavior that leads to confinement, is that false negatives can be

calculated and used to adjust the prediction method accordingly but false positives are indeterminable. As Foote (1970b) states, "if you lock a man up because you have diagnosed him to be dangerous, he has no chance to demonstrate that the diagnosis may have been mistaken" (p. 10). In short, accuracy can only be adjusted based on false negatives and adjusting to account for them will most likely be at the cost of many more false positives.

This thesis has tried to provide some cautions regarding predictions of dangerous and violent behavior; as Slobogin (2006) states, "It may be that abolishing dangerousness as a legal criterion would be the least costly approach to the problem" (p. 173). We are not advocating abolishing prediction methods for predicting future behavior, and certainly actuarial predictions should be used over clinical ones. What is being advocated is for a more objective evaluation of such predictions. The point is not that the methods should be abolished, but that uncertainty associated with them should be stated more clearly and that they may not be as good as advertised. There is a Latin phrase, *primum non nocere*, meaning "first, do no harm" and related to the Hippocratic Oath taken by healthcare professionals. The popular statistician Nate Silver states, "If you can't make a good prediction, it is very often harmful to pretend that you can" (N. Silver, 2012, p. 230). This philosophy should be considered whenever making decisions based on predictions of dangerous and violent behavior. The ultimate goal of risk assessment predictions has been to curtail crime, but when discussing preventive methods MacKenzie (2013) points out,

> [W]e have a social obligation not to harm either the individuals who come under the responsibility of the justice system or the society from which they come. Our recent policies have come at a great financial cost and caused damage to individuals and communities. And what we have been doing is not effective in reducing crime!" (p. 3).

By saying that a test fails to outperform base rate prediction, we are not suggesting that only base rate prediction should be done. But, by failing to outperform base rate prediction, we know that more than half of the positive predictions will be incorrect. This is

the point that should be understood and conveyed in the judicial system, or in any other area where predictions are made. The implications, of course, differ under different contexts. In considering parole decisions where predictions involve the incarcerated population, predictions concern individuals serving time for a crime they were convicted of committing. In contrast, those on trial for a crime and subject to involuntary commitment, present a much different situation. As this thesis has shown, predicting future human behavior is difficult. Misleading researchers, academics, the legal system, and the general public with statistics that only tell a part of the story is ethically irresponsible. More work is necessary; we should definitely not simply settle with what is available because the AUC happens to be "significant."

It needs to be emphasized that more information is necessary to properly assess predictive measures, and simply providing an AUC is insufficient. At the least, the positive and negative predictive values should be included, possibly in the form of an AUC-like measure or figure (for example, see Frederick & Bowden, 2009). The positive predictive value should be considered the central measure when the base rate is less than one half. The Brier score (Brier, 1950) is yet another way to better portray the accuracy of a measure, particularly in its decomposed form. The use of contingency tables to display the results should be provided routinely, either in text or through supplemental material; it allows researchers to assess other accuracy measures apart from the AUC.

Failures to cross-validate are simply unacceptable. No prediction model deserves consideration when it has not been cross-validated—it is far too likely that the model overfits the data. With today's inexpensive computing power, there is no excuse for not cross-validating a prediction tool. When adjusting the cutscore of a prediction model, one needs to consider the implications; for example, when using classification trees, cutscore adjustment is equivalent to assigning differing costs to false negatives and false positives—when these costs are not justified, cutscore adjustment should not be done.

Although widely popular, the use of unbiased predictors is inappropriate when other

predictors with smaller mean squared error exist, even when they are biased. Biased predictors should always be given consideration whenever an improvement in prediction accuracy is possible.

The prediction of violence is controversial; although it may be possible to predict those more likely to commit acts of violence (i.e., the *violence prone*), those more likely to commit violence may not in fact, actually do so. Speculatively, it may be that these individuals need to be "set off" by an event beyond anyone's control, triggering a set of reactions that leads to violence. Predicting this trigger may be impossible, and the events leading to this point are part a dynamic system consisting of a series of unpredictable events.

The movie *Falling Down* starring Michael Douglas comes to mind. In *Falling Down*, Michael Douglass plays a "disturbed man" where "nothing seems to be going right"; he "finally snaps" and "reache[s] his boiling point" (Hartill, O'Cain, & Sutton, 2013). Of course this is an over-dramatized version of a "set off" effect, but if the Korean store owner had given him change for a dollar, would there have been a movie? The case in point is that although William Foster (Douglas's character) may have been someone who was violent prone, it took a series of events to elicit this otherwise latent violent behavior.

So the question becomes, do you lock someone up because they could potentially commit act(s) of violence? (Certainly there are alternatives with therapy being the first that comes to mind.) This question was dissected in an entertaining confabulation between the White Queen and Alice in Lewis Carrol's book, *Through the Looking-Glass, and What Alice Found There* (1875):

> "It's a poor sort of memory that only works backwards," the Queen remarked.
>
> "What sort of things do you remember best?" Alice ventured to ask.
>
> "Oh, things that happened the week after next," the Queen replied in a careless tone. "For instance, now," she went on, sticking a large piece of plaster on her finger as she spoke, "there's the King's Messenger. He's in prison now, being punished: and the trial doesn't even begin till next Wednesday: and of course the crime comes

last of all."

"Suppose he never commits the crime?" said Alice.

"That would be all the better, wouldn't it?" the Queen said, as she bound the plaster round her finger with a bit of ribbon.

Alice felt there was no denying that. "Of course it would be all the better," she said: "but it wouldn't be all the better his being punished."

"You're wrong there, at any rate," said the Queen: "were you ever punished?"

"Only for faults," said Alice.

"And you were all the better for it, I know!" the Queen said triumphantly.

"Yes, but then I had done the things I was punished for," said Alice: "that makes all the difference."

The well-respected statistician George Box provides the epigraph at the beginning of this chapter. A similar saying might apply to psychologists: psychologists, like models, have the bad habit of falling in love with their measurements.

Appendix A

Proofs

A.1 The Meehl-Rosen Condition

From Chapter 3.

Proof.

$$
\begin{aligned}
P(A) &> \frac{1 - P(\bar{B}|\bar{A})}{P(B|A) + (1 - P(\bar{B}|\bar{A}))} \\
P(A)[P(B|A) + (1 - P(\bar{B}|\bar{A}))] &> 1 - P(\bar{B}|\bar{A}) \\
P(B|A)P(A) + P(A) - P(\bar{B}|\bar{A})P(A) &> 1 - P(\bar{B}|\bar{A}) \\
P(B|A)P(A) - P(\bar{B}|\bar{A})P(A) + P(\bar{B}|\bar{A}) &> 1 - P(A) \\
P(B|A)P(A) + P(\bar{B}|\bar{A})(1 - P(A)) &> P(\bar{A}) \\
P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) &> P(\bar{A}).
\end{aligned}
$$

□

A.2 The Dawes Condition

From Chapter 3.

Before proving the Dawes condition, note that $P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$ and $P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) = P(A)$. These two equalities are used in the proof.

Proof.

$$
\begin{aligned}
P(\bar{A}|B) &< \frac{1}{2} \\
2P(\bar{A}|B) &< 1 \\
P(\bar{A}|B) + P(\bar{A}|B) &< 1 \\
P(\bar{A}|B) &< 1 - P(\bar{A}|B) \\
P(\bar{A}|B) &< P(A|B) \\
P(\bar{A}|B)P(B) &< P(A|B)P(B) \\
P(\bar{A}|B)P(B) + P(\bar{A}|\bar{B})P(\bar{B}) &< P(A|B)P(B) + P(\bar{A}|\bar{B})P(\bar{B}) \\
P(\bar{A}) &< P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A})
\end{aligned}
$$

□

A.3 The Bokhari-Hubert Condition

From Chapter 3.

Proof. We begin with the general condition:

$$
\begin{aligned}
P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) &> P(\bar{A}) \\
\frac{n_{BA} + n_{\bar{B}\bar{A}}}{n} &> \frac{n_{\bar{A}}}{n} \\
n_{BA} + n_{\bar{B}\bar{A}} &> n_{\bar{A}} \\
n_{BA} + n_{\bar{B}\bar{A}} &> n_{B\bar{A}} + n_{\bar{B}\bar{A}} \\
n_{BA} &> n_{B\bar{A}}.
\end{aligned}
$$

Thus, $P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) > P(\bar{A}) \Leftrightarrow n_{BA} > n_{B\bar{A}}$. Now,

$$\begin{aligned} n_{BA} &> n_{B\bar{A}} \\ n_{\bar{B}A} + n_{BA} + n_{\bar{B}\bar{A}} &> n_{\bar{B}A} + n_{B\bar{A}} + n_{\bar{B}\bar{A}} \\ n_{\bar{B}\bar{A}} &> n_{\bar{B}A} + n_{B\bar{A}} + n_{\bar{B}\bar{A}} - (n_{\bar{B}A} + n_{BA}) \\ n_{\bar{B}\bar{A}} &> n_{\bar{B}A} + n_{\bar{A}} - n_A. \end{aligned}$$

Because (by assumption) $n_{\bar{A}} > n_A$, we have $n_{\bar{A}} - n_A > 0$, and therefore, $n_{\bar{B}\bar{A}} > n_{\bar{B}A}$, as desired. □

A.4 The Bokhari-Hubert Condition and Relative Risk

From Chapter 3.

We assume the Bokhari-Hubert condition holds.

Proof.

$$\begin{aligned} n_{BA} > n_{B\bar{A}} \text{ and } n_{\bar{B}\bar{A}} > n_{\bar{B}A} &\Rightarrow n_{BA}n_{\bar{B}\bar{A}} > n_{B\bar{A}}n_{\bar{B}A} \\ n_{BA}n_{\bar{B}\bar{A}} + n_{BA}n_{\bar{B}A} &> n_{B\bar{A}}n_{\bar{B}A} + n_{BA}n_{\bar{B}A} \\ n_{BA}\left(n_{\bar{B}\bar{A}} + n_{\bar{B}A}\right) &> n_{\bar{B}A}\left(n_{B\bar{A}} + n_{BA}\right) \\ n_{BA}n_{\bar{B}} &> n_{\bar{B}A}n_B \\ \frac{n_{BA}n_{\bar{B}}}{n_{\bar{B}A}n_B} &> 1 \\ RR &> 1 \end{aligned}$$

□

A.5 The Area Under the Curve

From Chapter 4.

Proof. From Figure 4.1, the labels of A, B, and C conform to areas:

$$A = (1 - \text{sensitivity})(1 - \text{specificity});$$
$$B = \frac{1}{2}(1 - \text{specificity})(\text{sensitivity});$$

and

$$C = \frac{1}{2}(1 - \text{sensitivity})(\text{specificity}).$$

Thus, the area under the ROC curve is

$$\text{AUC} = 1 - (A + B + C) = \frac{\text{specificity} + \text{sensitivity}}{2}.$$

□

A.6 Relationship Between Dawes' Properties

From Chapter 4.

We show that Properties (2) and (3) imply Property (1).

Proof. By Property (3), $P(A) < 1/2$ and $P(B) < 1/2$. For some constant $0 < k < 1/2$, we have $P(A) = 1/2 - k$. Redefining $P(B) = P(B|A)P(A) + P(B|\bar{A})(1 - P(A))$ and substituting for $P(A)$, we have

$$P(B|A)P(A) + P(B|\bar{A})(1 - P(A)) < \frac{1}{2}$$

$$P(B|A)\left(\frac{1}{2}-k\right)+P(B|\bar{A})\left(1-\left(\frac{1}{2}-k\right)\right)<\frac{1}{2}$$
$$\left(\frac{1}{2}\right)\big(P(B|A)+P(B|\bar{A})\big)+k\big(P(B|\bar{A})-P(B|A)\big)<\frac{1}{2}.$$

By Property (2)

$$P(B|A)>P(\bar{B}|\bar{A})=1-P(B|\bar{A})\Leftrightarrow P(B|A)+P(\bar{B}|\bar{A})>1,$$

and because $k>0$, it must be true that $P(B|A)>P(\bar{B}|\bar{A})$, otherwise we contradict Property (3) (namely, $P(B)<1/2$). □

A.7 Reduction in Accuracy Measures

From Chapter 4.

Given that the three Dawes' properties hold, we show that Equation (4.3) is necessarily true.

Proof. We first show that AUC > Acc. Consider Property (2):

$$P(B|A)>P(\bar{B}|\bar{A}).$$

Because Property (3) states that $P(A)<1/2$, we have that $1-2P(A)>0$; thus,

$$\begin{aligned}
P(B|A)&>P(\bar{B}|\bar{A})\\
P(B|A)\big(1-2P(A)\big)&>P(\bar{B}|\bar{A})\big(1-2P(A)\big)\\
P(B|A)\big(1-2P(A)\big)&>P(\bar{B}|\bar{A})\big(2P(\bar{A})-1\big)\\
P(B|A)-2P(B|A)P(A)&>2P(\bar{B}|\bar{A})P(\bar{A})-P(\bar{B}|\bar{A})\\
P(B|A)+P(\bar{B}|\bar{A})&>2\big(P(B|A)P(A)+P(\bar{B}|\bar{A})P(\bar{A})\big)\\
\frac{P(B|A)+P(\bar{B}|\bar{A})}{2}&>P(A\cap B)+P(\bar{A}\cap\bar{B})
\end{aligned}$$

$$\text{AUC} > \text{Acc.}$$

Thus, Properties (2) and (3) imply the first inequality.

To establish the last inequality, we first demonstrate that $P(B) > P(A)$; consider Property (2):

$$\begin{aligned} P(B|A) &> P(\bar{B}|\bar{A}) \\ P(B|A) &> 1 - P(B|\bar{A}) \\ P(B|\bar{A}) &> 1 - P(B|A). \end{aligned}$$

Now because $P(A) < 1/2$ (by Property (3)), we have

$$\begin{aligned} P(B|\bar{A})P(\bar{A}) &> \big(1 - P(B|A)\big)P(A) \\ P(B|\bar{A})P(\bar{A}) + P(B|A)P(A) &> P(A) \\ P(B) &> P(A). \end{aligned}$$

Thus, Properties (2) and (3) imply that $P(\bar{A}) > P(\bar{B}) > P(B) > P(A)$. To show that the two properties imply the last inequality begin with

$$\begin{aligned} P(\bar{A}) &> P(B) \\ P(\bar{A})\big(P(\bar{B}|\bar{A}) + P(B|\bar{A})\big) &> P(B) \\ P(\bar{B}|\bar{A})P(\bar{A}) + P(B|\bar{A})P(\bar{A}) &> P(B|\bar{A})P(\bar{A}) + P(B|A)P(A) \\ P(\bar{B}|\bar{A})P(\bar{A}) &> P(B|A)P(A) \\ P(B|\bar{A})P(\bar{B}|\bar{A})P(\bar{A}) &> P(B|\bar{A})P(B|A)P(A) \\ P(B|\bar{A})P(\bar{B}|\bar{A})P(\bar{A}) &> \big(1 - P(\bar{B}|\bar{A})\big)P(B|A)P(A) \\ P(B|\bar{A})P(\bar{B}|\bar{A})P(\bar{A}) + P(\bar{B}|\bar{A})P(B|A)P(A) &> P(B|A)P(A) \end{aligned}$$

$$
\begin{aligned}
P(\bar{B}|\bar{A})\big(P(B|\bar{A})P(\bar{A}) + P(B|A)P(A)\big) &> P(B|A)P(A) \\
P(\bar{B}|\bar{A})P(B) &> P(B|A)P(A) \\
P(\bar{B}|\bar{A})P(B)P(\bar{A}) &> P(B|A)P(A)P(\bar{A}).
\end{aligned}
$$

Because $P(\bar{A}) > P(\bar{B})$ we have

$$
\begin{aligned}
P(\bar{B}|\bar{A})P(B)P(\bar{A}) &> P(B|A)P(A)P(\bar{B}) \\
\frac{P(\bar{B}|\bar{A})P(\bar{A})}{P(\bar{B})} &> \frac{P(B|A)P(A)}{P(B)}.
\end{aligned}
$$

Because $P(B) < 1/2$ we have that $1 - 2P(B) > 0$ so

$$
\begin{aligned}
P(\bar{B}|\bar{A})P(\bar{A})\left(\frac{1-2P(B)}{P(\bar{B})}\right) &> P(B|A)P(A)\left(\frac{1-2P(B)}{P(B)}\right) \\
P(\bar{B}|\bar{A})P(\bar{A})\left(\frac{2P(\bar{B})-1}{P(\bar{B})}\right) &> P(B|A)P(A)\left(\frac{1-2P(B)}{P(B)}\right) \\
P(\bar{B}|\bar{A})P(\bar{A})\left(2-\frac{1}{P(\bar{B})}\right) &> P(B|A)P(A)\left(\frac{1}{P(B)}-2\right) \\
2P(\bar{B}|\bar{A})P(\bar{A}) - \frac{P(\bar{B}|\bar{A})P(\bar{A})}{P(\bar{B})} &> \frac{P(B|A)P(A)}{P(B)} - 2P(B|A)P(A) \\
2\big(P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A})\big) &> \frac{P(B|A)P(A)}{P(B)} + \frac{P(\bar{B}|\bar{A})P(\bar{A})}{P(\bar{B})} \\
P(B \cap A) + P(\bar{B} \cap \bar{A}) &> \frac{P(A|B) + P(\bar{A}|\bar{B})}{2} \\
\text{Acc} &> \frac{\text{PPV} + \text{NPV}}{2}.
\end{aligned}
$$

□

A.8 The Bokhari-Hubert Condition: $P(A) = P(B)$

From Chapter 6.

Proof. Suppose $P(A) = P(B)$. Without loss of generality let $P(A) < P(\bar{A})$. Because $P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) > P(\bar{A}) \Leftrightarrow P(A|B)P(B) + P(\bar{A}|\bar{B})P(\bar{B}) > P(\bar{B})$, we have

$$\begin{aligned} P(A|B)P(B) + P(\bar{A}|\bar{B})P(\bar{B}) &> P(\bar{B}) \\ \frac{n_{BA}}{n_B}\left(\frac{n_B}{n}\right) + \frac{n_{\bar{B}\bar{A}}}{n_{\bar{B}}}\left(\frac{n_{\bar{B}}}{n}\right) &> \frac{n_{\bar{B}}}{n} \\ n_{BA} + n_{\bar{B}\bar{A}} &> n_{\bar{B}} \\ n_{BA} + n_{\bar{B}\bar{A}} &> n_{\bar{B}A} + n_{\bar{B}\bar{A}} \\ n_{BA} &> n_{\bar{B}A}. \end{aligned}$$

Because $n_{\bar{B}} > n_B$,

$$\begin{aligned} n_{BA} + n_{\bar{B}\bar{A}} &> n_{\bar{B}A} + n_{\bar{B}\bar{A}} \\ n_{B\bar{A}} + n_{BA} + n_{\bar{B}\bar{A}} &> n_{B\bar{A}} + n_{\bar{B}A} + n_{\bar{B}\bar{A}} \\ n_{\bar{B}\bar{A}} &> n_{\bar{B}A} + n_{\bar{B}A} + n_{\bar{B}\bar{A}} - (n_{B\bar{A}} + n_{BA}) \\ n_{\bar{B}\bar{A}} &> n_{\bar{B}A} + n_{\bar{B}} - n_B, \end{aligned}$$

and $n_{\bar{B}\bar{A}} > n_{\bar{B}A}$, as desired. □

A.9 The Bokhari-Hubert Condition and Consistency

From Chapter 6.

We assume that $P(A) = P(B)$. If the Bokhari-Hubert condition holds, then the consistencies of both a positive and negative decision is at least $1/3$.

Proof. For the consistency of positive decisions,

$$P(A \cap B|A \cup B) = \frac{n_{BA}}{n_{BA} + n_{B\bar{A}} + n_{B\bar{A}}} > \frac{n_{BA}}{n_{BA} + n_{BA} + n_{BA}} = \frac{1}{3};$$

similarly for the consistency of negative decisions,

$$P(\bar{A} \cap \bar{B} | \bar{A} \cup \bar{B}) = \frac{n_{\bar{B}\bar{A}}}{n_{\bar{B}\bar{A}} + n_{B\bar{A}} + n_{B\bar{A}}} > \frac{n_{\bar{B}\bar{A}}}{n_{\bar{B}\bar{A}} + n_{\bar{B}\bar{A}} + n_{\bar{B}\bar{A}}} = \frac{1}{3}.$$

□

A.10 The Bokhari-Hubert Condition and Diagnostic Likelihood Ratios: $P(A) = P(B)$

From Chapter 6. Suppose $P(A) = P(B)$. For this proof, we first note that

$$P(A) = P(B) \Leftrightarrow n_A = n_B \Leftrightarrow n_{BA} + n_{\bar{B}A} = n_{BA} + n_{B\bar{A}} \Leftrightarrow n_{\bar{B}A} = n_{B\bar{A}}.$$

We saw in Chapter 6 that, given $P(A) = P(B)$, if the Bokhari-Hubert condition holds, $n_{BA} > n_{\bar{B}A}$ and $n_{\bar{B}\bar{A}} > n_{B\bar{A}}$ (i.e., differential prediction also exists between the columns). We are ready to prove that, if $P(A) = P(B)$ and the Bokhari-Hubert condition is met, then the positive diagnostic likelihood ratio is greater than one and the negative diagnostic likelihood ratio is less than one.

Proof. We saw in Chapter 3 that if ${}^{n_{BA}}/_{n_{B\bar{A}}} > {}^{n_A}/_{n_{\bar{A}}}$, the positive diagnostic likelihood ratio is greater than one. Similarly, if ${}^{n_{\bar{B}A}}/_{n_{\bar{B}\bar{A}}} < {}^{n_A}/_{n_{\bar{A}}}$, the negative diagnostic likelihood ratio is less than one. For the positive diagnostic likelihood ratio, we have

$$\begin{aligned} \frac{n_{BA}}{n_{B\bar{A}}} &= \frac{2n_{BA}}{2n_{B\bar{A}}} \\ &> \frac{n_{BA} + n_{\bar{B}A}}{n_{\bar{B}\bar{A}} + n_{B\bar{A}}} \\ &= \frac{n_A}{n_{\bar{A}}}. \end{aligned}$$

Similarly, for the negative diagnostic likelihood ratio,

$$\begin{aligned}\frac{n_{\bar{B}A}}{n_{\bar{B}\bar{A}}} &= \frac{2n_{\bar{B}A}}{2n_{\bar{B}\bar{A}}} \\ &< \frac{n_{\bar{B}A} + n_{BA}}{n_{\bar{B}\bar{A}} + n_{B\bar{A}}} \\ &= \frac{n_A}{n_{\bar{A}}}.\end{aligned}$$

Thus, $DLR_B > 1$ *and* $DLR_{\bar{B}} < 1$, necessarily, if $P(A) = P(B)$ and the Bokhari-Hubert condition holds. □

A.11 $P(A)$ given fixed PPV and NPV

From Chapter 6. Suppose that both the PPV and NPV are fixed. We exclude the trivial cases when $1 - \text{NPV,PPV} = 0, 1$; that is, assume $n_{BA}, n_{\bar{B}A}, n_{B\bar{A}}, n_{\bar{B}\bar{A}} \neq 0$. We show that for $0 < 1 - \text{NPV} < \text{PPV} < 1$,

$$1 - \text{NPV} < P(A) < \text{PPV};$$

similarly, for $0 < \text{PPV} < 1 - \text{NPV} < 1$

$$\text{PPV} < P(A) < 1 - \text{NPV}.$$

Proof. We have, by definition,

$$\begin{aligned}&P(A)P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) \\ &P(A) = \text{PPV} \cdot P(B) + (1 - \text{NPV})(1 - P(B)) \\ &P(A) = P(B)(\text{PPV} + \text{NPV} - 1) + (1 - \text{NPV}).\end{aligned}$$

Solving for $P(B)$, we have

$$P(B) = \frac{P(A) + \text{NPV} - 1}{\text{PPV} + \text{NPV} - 1}.$$

By definition, $0 < P(B) < 1$. Let $1 - \text{NPV} < \text{PPV}$ so that $\text{PPV} + \text{NPV} - 1 > 0$. Thus,

$$\begin{aligned} 0 < \frac{P(A) + \text{NPV} - 1}{\text{PPV} + \text{NPV} - 1} < 1 \\ 0 < P(A) + \text{NPV} - 1 < \text{PPV} + \text{NPV} - 1 \\ 1 - \text{NPV} < P(A) < \text{PPV}. \end{aligned}$$

Similarly, if we let $\text{PPV} < 1 - \text{NPV}$ we have $\text{PPV} + \text{NPV} - 1 < 0$; thus,

$$\begin{aligned} 0 < \frac{P(A) + \text{NPV} - 1}{\text{PPV} + \text{NPV} - 1} < 1 \\ \text{PPV} + \text{NPV} - 1 < P(A) + \text{NPV} - 1 < 0 \\ \text{PPV} < P(A) < 1 - \text{NPV}. \end{aligned}$$

□

A.12 Lower Limit for NPV

From Chapter 6. Suppose at a given cutscore that both the sensitivity and specificity of a test are greater than .5; in other words, the point on the ROC curve for the given cutscore is above the line of discrimination. Now suppose that $P(A) < 1/2$. We show that the negative predictive value is greater than $1/2$.

Proof. First, define the negative predictive value using Bayes' Theorem:

$$\text{NPV} = P(\bar{A}|\bar{B}) = \frac{P(\bar{B}|\bar{A})P(\bar{A})}{P(\bar{B})}.$$

Note that $P(\bar{B}) = P(\bar{B}|\bar{A})P(\bar{A}) + P(\bar{B}|A)P(A)$, where $.5 < P(\bar{B}|\bar{A}), P(\bar{A}) \leq 1$ and $0 \leq$

$P(\bar{B}|A), P(A) < .5$ so $P(\bar{B}|\bar{A})P(\bar{A}) > P(\bar{B}|A)P(A)$. This gives the desired inequality:

$$\begin{aligned}\text{NPV} &= \frac{P(\bar{B}|\bar{A})P(\bar{A})}{P(\bar{B}|\bar{A})P(\bar{A}) + P(\bar{B}|A)P(A)} \\ &> \frac{P(\bar{B}|\bar{A})P(\bar{A})}{2P(\bar{B}|\bar{A})P(\bar{A})} \\ &= \frac{1}{2}.\end{aligned}$$

Similarly, if both the sensitivity and specificity of a test are greater than .5 and $P(A) > 1/2$, the PPV $> 1/2$ □

A.13 Bias-Variance Decomposition

From Chapter 6.

Proof. Note that all expectations and variances are conditioned on the data, X. Consider the risk function for the squared error loss:

$$\mathcal{R}(y, \hat{y}) = \mathbb{E}\left(\mathcal{L}_2(y, \hat{y})\right) = \mathbb{E}\left[(y - \hat{y})^2\right].$$

The risk function can be decomposed as follows:

$$\mathbb{E}\left[(y - f(x) + f(x) - \hat{y})^2\right] = \mathbb{E}\left[(y - f(x))^2\right] + \mathbb{E}\left[(f(x) - \hat{y})^2\right] + 2\mathbb{E}\left[(y - f(x))(f(x) - \hat{y})\right].$$

The first term reduces to

$$\mathbb{E}\left[(y - f(x))^2\right] = \mathbb{E}\left[(f(x) + \boldsymbol{\varepsilon} - f(x))^2\right] = \mathbb{E}\left(\boldsymbol{\varepsilon}^2\right) = \mathbb{V}(\boldsymbol{\varepsilon}).$$

The last term (disregarding the 2) reduces to

$$\mathbb{E}\left[(y - f(x))(f(x) - \hat{y})\right] = f(x)\mathbb{E}(y) - f(x)^2 + \mathbb{E}(y\hat{y}) - f(x)\mathbb{E}(\hat{y})$$

$$
\begin{aligned}
&= f(x)^2 - f(x)^2 + \mathbb{E}\left((f(x) + \varepsilon)\hat{y}\right) - f(x)\mathbb{E}(\hat{y}) \\
&= \mathbb{E}\left(f(x)\hat{y} + \varepsilon\hat{y}\right)) - f(x)\mathbb{E}(\hat{y}) \\
&= f(x)\mathbb{E}(\hat{y}) - f(x)\mathbb{E}(\hat{y}) = 0.
\end{aligned}
$$

Thus,

$$
\mathcal{R}(y, \hat{y}) = \mathbb{E}\left[(f(x) - \hat{y})^2\right] + \mathbb{V}(\varepsilon).
$$

The first term can be further decomposed as follows:

$$
\begin{aligned}
\mathbb{E}\left[(f(x) - \hat{y})^2\right] &= \mathbb{E}\left[(f(x) - \mathbb{E}(\hat{y}) + \mathbb{E}(\hat{y}) - \hat{y})^2\right] \\
&= \mathbb{E}\left[(f(x) - \mathbb{E}(\hat{y}))^2\right] + \mathbb{E}\left[(\mathbb{E}(\hat{y}) - \hat{y})^2\right] + 2\mathbb{E}\left[(f(x) - \mathbb{E}(\hat{y}))(\mathbb{E}(\hat{y}) - \hat{y})\right] \\
&= \mathbb{E}\left[\text{bias}(\hat{y})^2\right] + \mathbb{V}(\hat{y}) + 2\left[\mathbb{E}\left(f(x)\mathbb{E}(\hat{y})\right) - \mathbb{E}(f(x)\hat{y}) - \mathbb{E}(\mathbb{E}(\hat{y})\mathbb{E}(\hat{y})) + \mathbb{E}\left(\mathbb{E}(\hat{y})\hat{y}\right)\right] \\
&= \mathbb{E}\left[\text{bias}(\hat{y})^2\right] + \mathbb{V}(\hat{y}) + 2\left[f(x)\mathbb{E}\left(\hat{y}\right) - f(x)\mathbb{E}(\hat{y}) - \mathbb{E}(\hat{y})^2 + \mathbb{E}(\hat{y})^2\right] \\
&= \text{bias}(\hat{y})^2 + \mathbb{V}(\hat{y}).
\end{aligned}
$$

□

A.14 Brier Score Decomposition

From Chapter 6.

Proof. The Brier score for a given subset of estimated probabilities is given as

$$
\text{VS}^{(r)} = \frac{1}{n_r} \sum_{j=1}^{n_r} \|\mathbf{y}_j^{(r)} - \hat{\mathbf{y}}^{(r)}\|^2.
$$

This is decomposed as follows:

$$\begin{aligned}
\mathrm{VS}^{(r)} &= \frac{1}{n_r}\sum_{j=1}^{n_r}\left[\left(\mathbf{y}_j^{(r)}-\hat{\mathbf{y}}^{(r)}\right)^T\left(\mathbf{y}_j^{(r)}-\hat{\mathbf{y}}^{(r)}\right)\right] \\
&= \frac{1}{n_r}\sum_{j=1}^{n_r}\left(\mathbf{y}_j^{(r)T}\mathbf{y}_j^{(r)}-2\hat{\mathbf{y}}^{(r)T}\mathbf{y}_j^{(r)}+\hat{\mathbf{y}}^{(r)T}\hat{\mathbf{y}}^{(r)}\right) \\
&= \frac{1}{n_r}\sum_{j=1}^{n_r}\mathbf{y}_j^{(r)T}\mathbf{y}_j^{(r)}-2\hat{\mathbf{y}}^{(r)T}\frac{1}{n_r}\sum_{j=1}^{n_r}\mathbf{y}_j^{(r)}+\hat{\mathbf{y}}^{(r)T}\hat{\mathbf{y}}^{(r)} \\
&= \frac{1}{n_r}\sum_{j=1}^{n_r}\mathbf{y}_j^{(r)T}\mathbf{e}-2\hat{\mathbf{y}}^{(r)T}\frac{1}{n_r}\sum_{j=1}^{n_r}\mathbf{y}_j^{(r)}+\hat{\mathbf{y}}^{(r)T}\hat{\mathbf{y}}^{(r)},
\end{aligned}$$

where $\mathbf{e}$ is an $n_r \times 1$ vector of 1's. Let

$$\bar{\mathbf{y}}^{(r)} = \frac{1}{n_r}\sum_{j=1}^{n_r}\mathbf{y}_j^{(r)};$$

that is, $\bar{\mathbf{y}}^{(r)} = (p_1^{(r)}, p_2^{(r)})^T$ where $p_2^{(r)} = 1 - p_1^{(r)}$ is the base rate probability within the subset (i.e., the probability of the event occurring given the estimated probability).

Substituting $\bar{\mathbf{y}}^{(r)}$ into the previous equation gives

$$\begin{aligned}
\bar{\mathbf{y}}^{(r)T}\mathbf{e}-2\hat{\mathbf{y}}^{(r)T}\bar{\mathbf{y}}^{(r)}+\hat{\mathbf{y}}^{(r)T}\hat{\mathbf{y}}^{(r)} &= \bar{\mathbf{y}}^{(r)T}\mathbf{e}-2\hat{\mathbf{y}}^{(r)T}\bar{\mathbf{y}}^{(r)}+\hat{\mathbf{y}}^{(r)T}\hat{\mathbf{y}}^{(r)}+\bar{\mathbf{y}}^{(r)T}\bar{\mathbf{y}}^{(r)}-\bar{\mathbf{y}}^{(r)T}\bar{\mathbf{y}}^{(r)} \\
&= (\hat{\mathbf{y}}^{(r)}-\bar{\mathbf{y}}^{(r)})^T(\hat{\mathbf{y}}^{(r)}-\bar{\mathbf{y}}^{(r)})+\bar{\mathbf{y}}^{(r)T}(\mathbf{e}-\bar{\mathbf{y}}^{(r)}) \\
&= \|\hat{\mathbf{y}}^{(r)}-\bar{\mathbf{y}}^{(r)}\|^2+\langle\bar{\mathbf{y}}^{(r)},\mathbf{e}-\bar{\mathbf{y}}^{(r)}\rangle.
\end{aligned}$$

The overall Brier score is a weighted sum of the subsets:

$$\frac{1}{n}\sum_{r=1}^{R}n_r\|\hat{\mathbf{y}}^{(r)}-\bar{\mathbf{y}}^{(r)}\|^2+\frac{1}{n}\sum_{r=1}^{R}n_r\langle\bar{\mathbf{y}}^{(r)},\mathbf{e}-\bar{\mathbf{y}}^{(r)}\rangle. \tag{A.1}$$

This second term can be further decomposed. To show this, first note that $\sum_{r=1}^{R} n_r = n$ and

$$\frac{1}{n}\sum_{r=1}^{R} n_r \bar{\mathbf{y}}^{(r)} = \frac{1}{n}\sum_{r=1}^{R} n_r \left(\frac{1}{n_r}\sum_{j=1}^{n_r} \mathbf{y}_j^{(r)}\right) = \frac{1}{n}\sum_{r=1}^{R}\sum_{j=1}^{n_r} \mathbf{y}_j^{(r)} \equiv \bar{\mathbf{y}},$$

the vector containing the overall base rate probabilities, p_1 and p_2. Now, working with the second partition from Equation (A.1),

$$\begin{aligned}
\frac{1}{n}\sum_{r=1}^{R} n_r \bar{\mathbf{y}}^{(r)T}(\mathbf{e} - \bar{\mathbf{y}}^{(r)}) &= \frac{1}{n}\sum_{r=1}^{R} n_r \bar{\mathbf{y}}^{(r)T}\mathbf{e} - \frac{1}{n}\sum_{r=1}^{R} n_r \bar{\mathbf{y}}^{(r)T}\bar{\mathbf{y}}^{(r)} \\
&= \bar{\mathbf{y}}^T\mathbf{e} - \frac{1}{n}\sum_{r=1}^{R} n_r \bar{\mathbf{y}}^{(r)T}\bar{\mathbf{y}}^{(r)} \\
&= \bar{\mathbf{y}}^T\mathbf{e} - 2\bar{\mathbf{y}}^T\bar{\mathbf{y}} + 2\bar{\mathbf{y}}^T\bar{\mathbf{y}} - \frac{1}{n}\sum_{r=1}^{R} n_r \bar{\mathbf{y}}^{(r)T}\bar{\mathbf{y}}^{(r)} \\
&= \bar{\mathbf{y}}^T(\mathbf{e} - \bar{\mathbf{y}}) - \left(\frac{1}{n}\sum_{r=1}^{R} n_r \bar{\mathbf{y}}^{(r)T}\bar{\mathbf{y}}^{(r)} - 2\bar{\mathbf{y}}^T\bar{\mathbf{y}} + \bar{\mathbf{y}}^T\bar{\mathbf{y}}\right) \\
&= \bar{\mathbf{y}}^T(\mathbf{e} - \bar{\mathbf{y}}) - \frac{1}{n}\sum_{r=1}^{R} n_r \left(\bar{\mathbf{y}}^{(r)T}\bar{\mathbf{y}}^{(r)} - 2\bar{\mathbf{y}}^{(r)T}\bar{\mathbf{y}} + \bar{\mathbf{y}}^T\bar{\mathbf{y}}\right) \\
&= \bar{\mathbf{y}}^T(\mathbf{e} - \bar{\mathbf{y}}) - \frac{1}{n}\sum_{r=1}^{R} n_r \left(\bar{\mathbf{y}}^{(r)} - \bar{\mathbf{y}}\right)^T \left(\bar{\mathbf{y}}^{(r)} - \bar{\mathbf{y}}\right) \\
&= \langle \bar{\mathbf{y}}, \mathbf{e} - \bar{\mathbf{y}} \rangle - \frac{1}{n}\sum_{r=1}^{R} n_r \|\bar{\mathbf{y}}^{(r)} - \bar{\mathbf{y}}\|^2.
\end{aligned}$$

Substituting this into Equation (A.1) gives the new decomposition:

$$\frac{1}{n}\sum_{r=1}^{R} n_r \|\hat{\mathbf{y}}^{(r)} - \bar{\mathbf{y}}^{(r)}\|^2 - \frac{1}{n}\sum_{r=1}^{R} n_r \|\bar{\mathbf{y}}^{(r)} - \bar{\mathbf{y}}\|^2 + \langle \bar{\mathbf{y}}, \mathbf{e} - \bar{\mathbf{y}} \rangle.$$

□

Appendix B

Assessment Tools

Risk Factor	Codes		Scores
Age	18–24		1
	25 or older		0
Single	Lived with lover		
	for at least two years?		
	No		1
	Yes		0
Index non-sexual	Yes		1
violence	No		0
Prior non-sexual	Yes		1
violence	No		0
Prior sex	Charges	Convictions	
offenses	6+	4+	3
	3–5	2–3	2
	1–1	1	1
	0	0	0
Prior sentencing dates	4 or more		1
(excluding index)	3 or less		0
Any convictions for	Yes		1
non-contact sex offenses	No		0
Any unrelated victims	Yes		1
	No		0
Any stranger victims	Yes		1
	No		0
Any male victims	Yes		1
	No		0

Table B.1: Ten risk factors for the Static-99 and their scoring.

Risk Factor	**Codes**	**Scores**
Age at	18–24.9	3
release	25–34.9	2
	35–49.9	1
	50+	0
Persistence of sexual offending		
Sentencing occasions	4 or more	3
for sexual offenses	2–3	2
	1	1
	0	0
Juvenile arrest for sexual	Yes	1
offense (and convicted as	No	0
an adult for a separate offense)		
High rate of	rate greater than once every 15 years	1
sexual offending	rate less than once 15 years	0
Persistence	4–5	3
subscore	2–3	2
	1	1
	0	0
Deviant sexual interests		
Any convictions for	Yes	1
non-contact sex offenses	No	0
Male victims	Yes	1
	No	0
Two or more victims	Yes	1
$<$ 12 years, one unrelated	No	0
Deviant	3	3
subscore	2	2
	1	1

Table B.2 continued on next page.

Table B.2 (Cont.)

Risk Factor	**Codes**	**Scores**	
	0		0
Relationship to victims			
Any unrelated victims	Yes	1	
	No	0	
Any stranger victims	Yes	1	
	No	0	
Relationship	2		2
subscore	1		1
	0		0
General Criminality			
Arrest/sentencing occasions	14 or more sentencing occasions	3	
	3–13 prior sentencing occasions	2	
	any prior charges/convictions but less than 3 prior sentencing occasions	1	
	no prior charges for anything	0	
Any breach of conditional	Yes	1	
release	No	0	
Years free prior to	less than 4 years	1	
index offense	4 or more years	0	
Any convictions for	Yes	1	
non-sexual violence	No	0	
General criminality	5–6		3
subscore	3–4		2
	1–2		1
	0		0

Table B.2: Risk factors for the Static-2002 and their scoring.

Appendix C

Analysis Details

C.1 Variables

From Chapter 5.

Table C.1 displays the variables included in the initial analyses. The first column is a brief description; the second column is the coding used in the text; the third column consists of the variable codes used in the original dataset. All variables come from the SPSS file `baseline.sav` except `F12VIOL` and `PCLTOT`; they were from the SPSS file `follow_up_subjects.sav`.

Variable Description	Variable Coding	Original Variables Used
Response Variable		
Violence	`Violence`	`f12viol`
Predictor Variables		
Age	`Age`	`AGE`
BIS Non-Planning Subscale	`BISnp`	`BISPLN`
BPRS Activation Subscale	`BPRSa`	`OACTV`
BPRS Hostile-Suspiciousness Subscale	`BPRSh`	`OHOST`
BPRS Total Score	`BPRSt`	`OBPRS`
Child Abuse Seriousness	`ChildAbuse`	`Q5.5.1`, `Q5.5.2`, `Q5.5.3`, `Q5.5.4`, `Q5.5.5`, `Q5.5.6`
Loss of Consciousness (Head Injury)	`Consc`	`NEU2A`

Table C.1 continued on next page

Table C.1 (Cont.)

Variable Description	Variable Coding	Original Variables Used
Father Arrest History	DadArr	Q5.20A, Q5.20B
Father's Drug Use	DadDrug	Q5.19A, Q5.19B
Drug Abuse (DSM-III-R)	DrugAbuse	DSM16A, DSM16B, DSM17A, DSM17B
Employed Prior to Hospitalization	Emp	Q4.4
Violent Fantasies: Escalating Seriousness	FantEsc	Q7.1, Q7.7
Violent Fantasies: Single Target Focus	FantSing	Q7.1, Q7.6
Violent Fantasies: Target Present	FantTarg	Q7.1, Q7.8
Level of Functioning	Function	Q9.1, Q9.2, Q9.3, Q9.4, Q9.5, Q9.6
Grandiose Delusions	GranDel	DEL03.1
Previous Head Injuries	HeadInj	NEU4B.1, NEU4B.2, NEU4B.3, NEU4B.4, NEU4B.5, NEU4B.6, NEU4B.7, NEU4B.8, NEU4B.9
Legal Status for Hospitalization	LegalStatus	LEGALR
Novaco Anger Scale Behavioral Subscale	NASb	NASBEH
Number of Negative Relationships	NegRel	Q10.10N, Q10.11N, Q10.12N, Q10.13N
Psychopathy Checklist: Screening Version	PCL	PCLTOT
MacArthur Perceived Coercion Scale	PCS	Q1.8, Q1.11, Q1.14, Q1.21, Q1.22
Prior Arrest Frequency	PriorArr	FREQARR
Property Crime Arrest	PropCrime	PROPARR
Violent Before Hospitalized	RecViol2	VIOL
Schizophrenic	Schiz	DSM2A, DSM5A
Proportion of Social Network are Mental Health Professionals	SNMHP	SNMHP
Substance Abuse	SubAbuse	DSM14A, DSM14B, DSM15A DSM15B

Table C.1 continued on next page

Table C.1 (Cont.)

Variable Description	Variable Coding	Original Variables Used
		DSM16A, DSM16B, DSM17A, DSM17B
Admission Reason: Suicide	Suicide	QREAS.02
Threat/Control Override Symptoms	tco	K1.1, K1.3, K2.1, K2.3, K3.1, K3.3, K4.1, K4.3, K8.1, K8.3, K9.1, K9.3, K10.1, K10.3, K12.1, K12.3
Threats at Admission	Threats	QREAS.20, QREAS.21

Table C.1: Variables used in analyses, from the MacArthur Violence Risk Assessment Study Monahan et al. (2001).

What follows is a description based on those in Monahan et al. (2001), of the variables used in our analyses.

`Violence`

Binary variable indicating violence at first or second follow-up time

0 = No Violence, 1 = Violence

`Age`

Patient's age in years

`BISnp`

Barratt Impulsiveness Scale (BIS), non-planning subscale

Possible Range: 0–44

`BPRSa`

Brief Psychiatric Rating Scale (BPRS), activation rating

Possible Range: 3–21

`BPRSh`

Brief Psychiatric Rating Scale (BPRS), hostility rating

Possible Range: 3–21

`BPRSt`

Brief Psychiatric Rating Scale (BPRS), total score

Possible Range: 18–126

`ChildAbuse`

Patient's self-reported seriousness of child abuse

0 = none, 1 = bare hand only, with no physical injury

2 = with an object, with no physical injury, 3 = resulting in physical injury

`Consc`

Patient's self-reported prior loss of consciousness due to head injury

0 = No, 1 = Yes

`DadArr`

Patient's self-reported arrest history of father

0 = Never, 1 = At least once

`DadDrug`

Patient's self-reported excessive drug use of father

0 = less often, 1 = Weekly/daily

`DrugAbuse`

Drug abuse diagnosis by trained clinician using DSM-III-R Checklist

0 = No diagnosis, 1 = Diagnosis

`Emp`

Patient's self-reported employment status (within the last two months prior to hospitalization)

0 = Not employed, 1 = Employed

`FantEsc`

Patient self-reported having daydreams or thoughts about physically hurting or injuring some other persons that have become more serious since they first began

0 = Less serious or the same, 1 = More serious

`FantSing`

Patient self-reported having daydreams or thoughts about physically hurting or injuring some other persons that have been about the same person

0 = Different persons, 1 = Same person

`FantTarg`

Patient self-reported having daydreams or thoughts about physically hurting or injuring some other persons while being with or watching that person

0 = No, 1 = Yes

`Function`

Patient's self-reported level of functioning based on difficulty of specific activities

`GranDel`

Trained clinician's rating of the presence of grandiose delusions

`HeadInj`

Patient's self-report of any head injury (with or without loss of consciousness)

`LegalStatus`

Hospital admission record of patient's admission status

0 = Voluntary, 1 = Involuntary

`NASb`

Novaco Anger Scale (NAS), behavioral rating

Possible Range: 16–48

`NegRel`

Average number of unique individuals patient named as involved in a negative relationship

`PCL`

Total score on Psychopathy Checklist: Screening Version

Possible Range: 0–24

`PCS`

MacArthur Perceived Coercion Scale (PCS)

`PriorArr`

Patient's self-reported number of arrests since age 15

0 = None, 1 = One, 2 = Two, 3 = Three or more

`PropCrime`

Police record indicating patient was arrested (since age 18) for a property crime

0 = No arrests, 1=At least one arrest

`RecViol2`

Patient's self-report of violence in the two months prior to hospitalization

0 = No violence, 1 = Violence

`Schiz`

Schizophrenia diagnosis by trained clinician using DSM-III-R Checklist

0 = No diagnosis, 1 = Diagnosis

`SNMHP`

Proportion of patient's social network that are mental health professionals

`SubAbuse`

Drug or alcohol abuse diagnosis by trained clinician using DSM-III-R Checklist

0 = No diagnosis, 1 = Diagnosis

`Suicide`

Hospital admission record of patient's admission status indicated suicide

0 = No, Yes

`tco`

Clinically validated affirmative answers to thought/control override symptoms

0 = Not present, 1 = Present

`Threats`

Presence of argumentativeness and threatening verbal statements by patient when admitted to hospital

0 = No, Yes

Below are the original MacArthur VRAS items used to create the preceding variables and their details based in the coding manuals (`final_bl_clinical.pdf`, `final_bl_research.pdf`, and `final_follow_subj.pdf`).

`F12VIOL`

Violence during follow-up one or follow-up two

0 = No Violence, 1 = Violence

`AGE`

Age in years

`BISPLN`

Non-Planning subscale (11 items) for 30-item Barratt Impulsiveness Scale (BIS)

Possible Range: 0–44

`DEL03.1`

Delusions characterized as grandiose 0 = Not checked, 1 = Checked

`DSM2A`

DSM-III-R diagnosis: Schizophrenia

1 = Absent, 3 = Present

`DSM5A`

DSM-III-R diagnosis: Schizoaffective Disorder

1 = Absent, 3 = Present

DSM14A

DSM-III-R diagnosis: Alcohol dependence (Current)

1 = Absent, 3 = Present

DSM14B

DSM-III-R diagnosis: Alcohol dependence (Lifetime)

1 = Absent, 3 = Present

DSM15A

DSM-III-R diagnosis: Alcohol abuse (Current)

1 = Absent, 3 = Present

DSM15B

DSM-III-R diagnosis: Alcohol abuse (Lifetime)

1 = Absent, 3 = Present

DSM16A

DSM-III-R diagnosis: Drug dependence (Current)

1 = Absent, 3 = Present

DSM16B

DSM-III-R diagnosis: Drug dependence (Lifetime)

1 = Absent, 3 = Present

DSM17A

DSM-III-R diagnosis: Drug abuse (Current)

1 = Absent, 3 = Present

DSM17A

DSM-III-R diagnosis: Drug abuse (Lifetime)

1 = Absent, 3 = Present

FREQARR

Prior arrests (Frequency)

K1.1

In the past two months, have you believed people were spying on you?

0 = No, 1 = Yes

K1.3

[If yes to K1.1, then clinician] Rates whether belief is delusional

0 = No, 1 = Possibly, 2 = Yes

K2.1

In the past two months, has there been a time when you believed people were following you?

0 = No, 1 = Yes

K2.3

[If yes to K1.1, then clinician] Rates whether belief is delusional

0 = No, 1 = Possibly, 2 = Yes

K3.1

In the past two months, have you believed that you were being secretly tested or experimented on?

0 = No, 1 = Yes

K3.3

[If yes to K1.1, then clinician] Rates whether belief is delusional

0 = No, 1 = Possibly, 2 = Yes

K4.1

In the past two months, have you believed that someone was plotting against you or trying to hurt you or poison you?

0 = No, 1 = Yes

K4.3

[If yes to K1.1, then clinician] Rates whether belief is delusional

0 = No, 1 = Possibly, 2 = Yes

K8.1

In the past two months, did you feel that you were under the control of some person, power or force, so that your actions and thoughts were not your own?

0 = No, 1 = Yes

K8.3

[If yes to K1.1, then clinician] Rates whether belief is delusional

0 = No, 1 = Possibly, 2 = Yes

K9.1

In the past two months, have you felt that strange thoughts or thoughts that were not your own were being put directly into your mind?

0 = No, 1 = Yes

K9.3

[If yes to K1.1, then clinician] Rates whether belief is delusional

0 = No, 1 = Possibly, 2 = Yes

K10.1

In the past two months, have you felt that someone or something could take or steal your thoughts out of your mind?

0 = No, 1 = Yes

K10.3

[If yes to K1.1, then clinician] Rates whether belief is delusional

0 = No, 1 = Possibly, 2 = Yes

K12.1

In the past two months, have you felt strange forces working on you, as if you were being hypnotized or magic was being performed on you, or you were being hit by x-rays or laser beams?

0 = No, 1 = Yes

K12.3

[If yes to K1.1, then clinician] Rates whether belief is delusional

0 = No, 1 = Possibly, 2 = Yes

LEGALR

Subject's legal status upon admission

0 = Voluntary, 1 = Involuntary

NASBEH

Behavioral subscale (16 items) for 48-item Novaco Anger Scale (NAS), Part A

Possible Range: 16–48

NEU2A

Have you ever been knocked out, knocked dizzy, passed out, fainted, or blacked out?

0 = No, 1 = Yes

NEU4B.1

Accident type: Bicycle. [If yes] Were you injured?

0 = No, 1 = Yes, no head injury, 2 = Yes, head injury

NEU4B.2

Accident type: Auto. [If yes] Were you injured?

0 = No, 1 = Yes, no head injury, 2 = Yes, head injury

NEU4B.3

Accident type: Motorcycle. [If yes] Were you injured?

0 = No, 1 = Yes, no head injury, 2 = Yes, head injury

NEU4B.4

Accident type: Fall. [If yes] Were you injured?

0 = No, 1 = Yes, no head injury, 2 = Yes, head injury

NEU4B.5

Accident type: Fall/Stairs. [If yes] Were you injured?

0 = No, 1 = Yes, no head injury, 2 = Yes, head injury

NEU4B.6

Accident type: Drowning. [If yes] Were you injured?

0 = No, 1 = Yes, no head injury, 2 = Yes, head injury

NEU4B.7

Accident type: Punched/Hit. [If yes] Were you injured?

0 = No, 1 = Yes, no head injury, 2 = Yes, head injury

NEU4B.8

Accident type: Sports. [If yes] Were you injured?

0 = No, 1 = Yes, no head injury, 2 = Yes, head injury

tttNEU4B.9

Accident type: Other. [If yes] Were you injured?

0 = No, 1 = Yes, no head injury, 2 = Yes, head injury

OACTV

Activation subscale (Tension, Mannerisms and posturing, Excitement) for 18-item Brief Psychiatric Rating Scale (BPRS)

Possible Range: 3–21

OHOST

Hostile-Suspiciousness subscale (Hostility, Suspiciousness, Uncooperativeness) for BPRS

Possible Range: 3–21

`OBPRS`

Total score on BPRS

Possible Range: 18–126

`PCLTOT`

Total score on 12-item Psychopathy Checklist: Screening Version

Possible Range: 0–24

`PROPARR`

Prior arrest for crimes against property

The following five questions are part of the Interpersonal Relations Scale – Abbreviated (IRS-A). They form the MacArthur Perceived Coercion Scale.

`Q1.8`

I felt free to do what I wanted about coming to the hospital

0 = False, 1 = True

`Q1.11`

I chose to come to the hospital 0 = False, 1 = True

`Q1.14`

It was my idea to come to the hospital 0 = False, 1 = True

`Q1.21`

I had a lot of control over whether I went to the hospital 0 = False, 1 = True

`Q1.22`

I had more influence than anyone else on whether I came into the hospital 0 = False, 1 = True

Q4.4

During the past two months [before you came to the hospital], did you work for pay either full-time or part-time?

1 = No, 2 = Yes, full-time, 3 = Yes, part-time

Q5.5.1

Thinking about when you were a child, (up to age 12), did your parents ever beat or really hurt you with bare hand/fist?

0 = Never, 1 = Once, 2 = Twice,

3 = Sometimes, 4 = Frequently, 5 = Most of the time

Q5.5.2

Thinking about when you were a child, (up to age 12), did your parents ever beat or hit you with something hard?

0 = Never, 1 = Once, 2 = Twice,

3 = Sometimes, 4 = Frequently, 5 = Most of the time

Q5.5.3

Thinking about when you were a child, (up to age 12), did your parents ever beat or hit you with a whip, strap, or belt?

0 = Never, 1 = Once, 2 = Twice,

3 = Sometimes, 4 = Frequently, 5 = Most of the time

Q5.5.4

Thinking about when you were a child, (up to age 12), did your parents ever injure you with a knife, gun, or other weapon?

0 = Never, 1 = Once, 2 = Twice,

3 = Sometimes, 4 = Frequently, 5 = Most of the time

Q5.5.5

Thinking about when you were a child, (up to age 12), did your parents ever hurt you badly enough that you needed a doctor?

0 = Never, 1 = Once, 2 = Twice,

3 = Sometimes, 4 = Frequently, 5 = Most of the time

Q5.5.6

Thinking about when you were a child, (up to age 12), did your parents ever physically injure you so that you were admitted to a hospital?

0 = Never, 1 = Once, 2 = Twice,

3 = Sometimes, 4 = Frequently, 5 = Most of the time

Q5.19A

Did your (biological) father ever use street drugs? If yes, How often?

0 = Never, 1 = Daily, 2 = Twice a week,

3 = Once a week, 4 = Once a month, 5 = Less often,

6 = Yes, but unknown frequency

Q5.19B

Did the man who raised you (not biological father) ever use street drugs? If yes, How often?

0 = Never, 1 = Daily, 2 = Twice a week,

3 = Once a week, 4 = Once a month, 5 = Less often,

6 = Yes, but unknown frequency

Q5.20A

Did your (biological) father ever get arrested? If yes, How many times?

0 = Never, 1 = 1 time, 2 = 2 times,

3 = 3 times, 4 = 4 times, 5 = 5–10 times,

6 = More than 10 times, 7 = More than once (unspecified)

Q5.20B

Did the man who raised you (not biological father) ever get arrested? If yes, How many times?

0 = Never, 1 = 1 time, 2 = 2 times,

3 = 3 times, 4 = 4 times, 5 = 5–10 times,

6 = More than 10 times, 7 = More than once (unspecified)

Q7.1

Do you ever have daydreams or thoughts about physically hurting or injuring some other persons?

1 = No, 2 = Yes

Q7.6

[If answered yes to Q7.1] Are they usually about the same person, or might they be about many different people?

1 = Same Person, 2 = Different people

Q7.7

[If answered yes to Q7.1] Since the time you first started having these thoughts, have the injuries that you think about gotten worse and worse, or have they always been about the same?

1 = Less Serious, 2 = Same, 3 = More Serious

Q7.8

[If answered yes to Q7.1] In the past two months, have you ever had these thoughts while actually being with or watching the person that you imagine hurting?

0 = No, 1 = Yes

Q9.1

How much difficulty do you usually have (or would you have) doing housework by yourself?

0 = None, 1 = Some, 2 = A lot, 3 = Unable to do it

Q9.2

How much difficulty do you usually have (or would you have) shopping for food or buying the things you usually need for yourself?

0 = None, 1 = Some, 2 = A lot, 3 = Unable to do it

Q9.3

How much difficulty do you usually have (or would you have) managing your money by yourself (such as keeping track of expenses, paying bills or making money last until the end of the month?)

0 = None, 1 = Some, 2 = A lot, 3 = Unable to do it

Q9.4

How much difficulty do you usually have using transportation?

0 = None, 1 = Some, 2 = A lot, 3 = Unable to do it

Q9.5

In the last two months, how much difficulty did you have (would you have) making your own meals or cooking for yourself on a regular basis?

0 = None, 1 = Some, 2 = A lot, 3 = Unable to do it

Q9.6

How much difficulty do you have (or would you have) doing laundry by yourself?

0 = None, 1 = Some, 2 = A lot, 3 = Unable to do it

Q10.10N

The number of persons mentioned to the question, Of the people you've mentioned [in your social network], is there anyone who asks so much of you that it bothers you?

Q10.11N

The number of persons mentioned to the question, Of the people you've mentioned [in your social network], when you go for help, does anyone turn you away?

Q10.12N

The number of persons mentioned to the question, Is there anyone you've mentioned [in your social network] with whom you really don't get along, or don't like, or who really upsets you?

Q10.13N

The number of persons mentioned to the question, Is there anyone you've mentioned [in your social network] who really doesn't seem to like you or who you seem to upset?

`QREAS.02`

Reason for current hospital admission: Suicide threat/Suicide ideation

0 = No, not checked, 1 = Checked

`QREAS.20`

Reason for current hospital admission: Argument/Threat

0 = No, not checked, 1 = Checked

`QREAS.21`

Reason for current hospital admission: Homicidal threat/Ideation

0 = No, not checked, 1 = Checked

`SNMHP`

Proportion of social network that are mental health professionals

`VIOL`

Violence (Previous two months)

0 = None, 2 = Violence

C.2 Software Code

From Chapter 5.

The SPSS file `baseline.sav` was imported into `R` and was combined with the variables `F12VIOL` and `PCLTOT` from the SPSS file `follow_up_subjects.sav` to create the dataset `COVR`.

C.2.1 Preprocessing the Data

The preprocessing code uses a user-defined function for determining the mode of a dataset:

```
varMode <- function(x) {
  varMode = as.numeric(names(table(x))[table(x) == max(table(x))])
}
```

The code that follows was used to preprocess the data so the variables match, as measured by matching correlations (see Table 5.1), the variables used in Monahan et al. (2001).

```
##### Response Variable
#violence committed? (0 = No, 1 = Yes)
Violence = as.factor(COVR$F12VIOL)

##### Predictor Variables
#Barratt Impulsiveness Scale (BIS) non-planning subscale [1,44]
BISnp = as.numeric(COVR$BISNPLN)
BISnp[is.na(BISnp)] = mean(na.omit(BISnp))

#brief psychiatric rating scale (BPRS) activation subscale [3,14]
BPRSa = as.numeric(COVR$OACTV)
BPRSa[is.na(BPRSa)] = mean(na.omit(BPRSa))

#BPRS hostile-suspiciousness subscale [3,18]
BPRSh = as.numeric(COVR$OHOST)
BPRSh[is.na(BPRSh)] = mean(na.omit(BPRSh))

#BPRS total score [18,74]
BPRSt = as.numeric(COVR$OBPRS)
BPRSt[is.na(BPRSt)] = mean(na.omit(BPRSt))

#Child Abuse Seriousness {0,1,2,3}
ChildAbuseVars = cbind(COVR$Q5.5.1,COVR$Q5.5.2,COVR$Q5.5.3,COVR$Q5.5.4,COVR$Q5.5.5,COVR$Q5.5.6)
  - 1
ChildAbuse = matrix(NA, 939, 1)
#0 = none
ChildAbuse0 = ChildAbuseVars
ChildAbuse0 = rowSums(ChildAbuse0)
ChildAbuse0[rowSums(is.na(ChildAbuseVars)) == 6] = NA
```

```
ChildAbuse[ChildAbuse0 == 0] = 0
rm(ChildAbuse0)
#1 = bare hand only, with no physical injury
#if Q5.5.1 > 0
ChildAbuse1 = ChildAbuseVars[,1]
ChildAbuse1[ChildAbuse1 > 1] = 1
#2 = with an object, with no physical injury
#if at least one of Q5.5.2, Q5.5.3 are > 0
ChildAbuse2 = ChildAbuseVars[,c(2:3)]
ChildAbuse2 = rowSums(ChildAbuse2)
ChildAbuse2[ChildAbuse2 > 1] = 1
#3 = resulting 1n phys1cal injury
#if at least one of Q5.5.4, Q5.5.5, Q5.5.6 are > 0
ChildAbuse3 = ChildAbuseVars[,c(4:6)]
ChildAbuse3 = rowSums(ChildAbuse3)
ChildAbuse3[ChildAbuse3 > 1] = 1
ChildAbuse[ChildAbuse1 > 0 & rowSums(cbind(ChildAbuse2, ChildAbuse3)) == 0,] = 1
ChildAbuse[ChildAbuse2 > 0 & ChildAbuse3 == 0,] = 2
ChildAbuse[ChildAbuse3 > 0,] = 3
rm(ChildAbuse1)
rm(ChildAbuse2)
rm(ChildAbuse3)
ChildAbuse[is.na(ChildAbuse)] = mean(ChildAbuse)
rm(ChildAbuseVars)

#employed prior to hospitalization (0 = no, 1 = FT, 2 = PT)
Emp = as.numeric(COVR$Q4.4) - 1 #"YES - FULL-TIME" = 1, "YES - PART-TIME" = 2, "NO" = 0
Emp[Emp == 2] = 1 #employed 1, unemployed 0
Emp[is.na(Emp)] = varMode(Emp)

#father's drug use (0 = less often, 1 = weekly/daily)
DadDrug = cbind(COVR$Q5.19A,COVR$Q5.19B) - 1
DadDrug[DadDrug > 6] = NA
DD = matrix(NA, 939, 2)
DD[DadDrug %in% 1:3] = 1
DD[DadDrug %in% c(0,4:6)] = 0
DD[is.na(DD)] = 0
DD = rowSums(DD)
DD[DD == 2] = 1
DD[rowSums(is.na(DadDrug)) == 2] = NA
DD[rowSums(DadDrug) == 0] = NA #if father never used drugs
```

```
DD[is.na(DD)] = varMode(DD)
DadDrug = DD
rm(DD)

#prior arrest history (frequency, 0 = none, 1 = one, 2 = two, 3 = 3+)
PriorArr = as.numeric(COVR$FREQARR) - 1
PriorArr[is.na(PriorArr)] = mean(na.omit(PriorArr))

#presence of grandiose delusions (0 = not present, 1 = present)
GranDel = as.numeric(COVR$DEL03.1) - 1
GranDel[is.na(GranDel)] = varMode(na.omit(GranDel))

#involuntary hospitalization admission status (0 = voluntary, 1 = involuntary)
LegalStatus = as.numeric(COVR$LEGALR) - 1
LegalStatus[is.na(LegalStatus)] = varMode(LegalStatus)

#proportion of social network members who are also mental health professionals [0,1]
SNMHP = as.numeric(COVR$SNMHP)
SNMHP[is.na(SNMHP)] = mean(na.omit(SNMHP))

#Novaco Anger Scale (NAS) behavioral subscale [16,48]
NASb = as.numeric(COVR$NOVBEH)
NASb[is.na(NASb)] = mean(na.omit(NASb))

#loss of consciousness due to head injury (0 = no, 1 = yes)
Consc = as.numeric(COVR$NEU2A) - 1
Consc[Consc > 1] = NA
Consc[Consc == 9] = as.factor(varMode(Consc))

#Psychopathy Checklist: Screening Version (PCL:SV) total score (0 = 0-12, 1 = 13-24)
PCL = as.numeric(COVR$PCLTOT)
PCL = round(PCL)
PCL[PCL < 13] = 0 #low
PCL[PCL > 12] = 1 #high
PCL[is.na(PCL)] = varMode(PCL)

#DSM-III-R Checklist: Drug abuse (0 = no, 1 = yes)
DrugAbuse = cbind(COVR$DSM16A,COVR$DSM16B,COVR$DSM17A,COVR$DSM17B) - 1
DrugAbuse[DrugAbuse == 5] = NA
da = DrugAbuse
da[is.na(da)] = 0
```

```
da = rowSums(da)
da[da > 1] = 1
da[rowSums(is.na(DrugAbuse)) == 4] = NA
DrugAbuse = da
DrugAbuse[is.na(DrugAbuse)] = varMode(DrugAbuse)
rm(da)

#threat/control override symptoms (0 = not present/verified, 1 = clinician verified)
tco.total = cbind(COVR$K1.1,COVR$K1.3,COVR$K2.1,COVR$K2.3,COVR$K3.1,COVR$K3.3,
  COVR$K4.1,COVR$K4.3,COVR$K8.1,COVR$K8.3,COVR$K9.1,COVR$K9.3,
  COVR$K10.1,COVR$K10.3,COVR$K12.1,COVR$K12.3) - 1
#patient's answers
tco.total[is.na(tco.total)] = 0
tco.patient = tco.total[,((1:8)*2-1)]
#clinician verification
tco.clinical = tco.total[,(1:8)*2]
tco.clinical[tco.clinical == 1] = 0
tco.clinical[tco.clinical == 2] = 1
#no individual refused to answer all 8 questions
tco = matrix(0, 939, 1)
#patient answered yes (1) and clinician verified (yes = 2)
tco[rowSums((tco.patient + tco.clinical) == 2) > 0] = 1
#no missing data
rm(tco.total,tco.patient,tco.clinical)

#violent fantasies with escalating seriousness (0 = no, 1 = yes)
viofan = cbind(COVR$Q7.1,COVR$Q7.7) - 1
viofan[viofan >= 3] = NA
viofan[viofan[,2] < 2,2] = 0 #less serious or the same
viofan[viofan[,2] == 2,2] = 1 #more serious
FantEsc = matrix(0, 939, 1)
FantEsc[is.na(viofan[,1])] = NA
FantEsc[viofan[,2] == 1] = 1
FantEsc[is.na(FantEsc)] = varMode(FantEsc)
rm(viofan)

#violent fantasies with single target focus (0 = no, 1 = yes)
viofan = cbind(COVR$Q7.1,COVR$Q7.6)
viofan[viofan >= 3] = NA
viofan[viofan[,2] == 2,2] = 0 #different person
FantSing = matrix(0, 939, 1)
```

```
FantSing[is.na(viofan[,1])] = NA
FantSing[viofan[,2] == 1] = 1
FantSing[is.na(FantSing)] = varMode(FantSing)
rm(viofan)

#self-reported violence two months prior to hospitalization (0 = no, 1 = yes)
RecViol2 = as.numeric(COVR$VIOL) - 1
#no missing data

#alcohol or drug abuse (0 = no, 1 = yes)
SubAbuse = cbind(COVR$DSM14A,COVR$DSM14B,COVR$DSM15A,COVR$DSM15B,
  COVR$DSM16A,COVR$DSM16B,COVR$DSM17A,COVR$DSM17B) - 1
SA = SubAbuse
SA[SA == 5] = NA
SubAbuse[SubAbuse == 5] = 0
SubAbuse = rowSums(SubAbuse)
SubAbuse[SubAbuse > 0] = 1
SubAbuse[rowSums(is.na(SA)) == 8] = NA
SubAbuse[is.na(SubAbuse)] = varMode(SubAbuse)
rm(SA)

#admission reason: suicide (0 = no, 1 = yes)
Suicide = cbind(COVR$QREAS.02) - 1
Suicide[is.na(Suicide)] = varMode(Suicide)

#father ever arrested (0 = no, 1 = yes)
DadArr = cbind(COVR$Q5.20A,COVR$Q5.20B) - 1
DadArr[DadArr == 8] = NA
DadArr[DadArr > 1] = 1
DadArr = cbind(DadArr,cbind(COVR$Q5.22A,COVR$Q5.22B) - 1)
DadArr[DadArr > 1] = NA
DA = DadArr
DadArr[is.na(DadArr)] = 0
DadArr = rowSums(DadArr)
DadArr[DadArr > 1] = 1
DadArr[rowSums(is.na(DA)) == 4] = NA
DadArr[is.na(DadArr)] = varMode(DadArr)
rm(DA)

#any previous head injury (0 = no, 1 = yes)
HeadInj = cbind(COVR$NEU4B.1,COVR$NEU4B.2,COVR$NEU4B.3,COVR$NEU4B.4,
```

```
  COVR$NEU4B.5,COVR$NEU4B.6,COVR$NEU4B.7,COVR$NEU4B.8,COVR$NEU4B.9) - 1
HeadInj[HeadInj == 3] = NA
HeadInj[HeadInj == 1] = 0
HeadInj[HeadInj == 2] = 1
HI = HeadInj
HeadInj[is.na(HeadInj)] = 0
HeadInj = rowSums(HeadInj)
HeadInj[HeadInj > 1] = 1
HeadInj[rowSums(is.na(HI)) == ncol(HI)] = NA
HeadInj[is.na(HeadInj)] = varMode(HeadInj)
rm(HI)

#violent fantasies with target was present (0 = not present, 1 = present)
viofan = cbind(COVR$Q7.1,COVR$Q7.8) - 1
viofan[viofan > 1] = NA
FantTarg = viofan
FantTarg[is.na(FantTarg)] = 0
FantTarg = rowSums(FantTarg)
FantTarg[FantTarg == 1] = 0
FantTarg[FantTarg == 2] = 1
FantTarg[rowSums(is.na(viofan)) == 2] = NA
FantTarg[is.na(FantTarg)] = varMode(FantTarg)
rm(viofan)

#diagnosis of schizophrenia (0 = no, 1 = yes)
Schiz = cbind(COVR$DSM2A,COVR$DSM5A) - 1
Schiz[Schiz == 5] = NA
Schiz[Schiz > 1] = 1
S = Schiz
Schiz[is.na(Schiz)] = 0
Schiz = rowSums(Schiz)
Schiz[Schiz > 1] = 1
Schiz[rowSums(is.na(S)) == 2] = NA
Schiz[is.na(Schiz)] = varMode(Schiz)
rm(S)

#age of patient [18,40]
Age = COVR$AGE
#no missing data

#level of functioning
```

```
#(6 items: 0 = None; 1 = Some; 2 = A lot; 3 = Unable to do it; Sum: [0,18])
Function = cbind(COVR$Q9.1,COVR$Q9.2,COVR$Q9.3,COVR$Q9.4,COVR$Q9.5,COVR$Q9.6) - 1
F = Function
Function[is.na(Function)] = 0
Function = rowSums(Function)
Function[rowSums(is.na(F)) == 5] = NA
Function[is.na(Function)] = mean(Function)
rm(F)

#arrested since age 18 for property crime (0 = no, 1 = yes)
PropCrime = as.numeric(COVR$PROPARR) - 1
#no missing data

#MacArthur Perceived Coercion Scale (PCS) (0 = True, 1 = False; Summed Score, 0-5)
PCS = -1*(cbind(COVR$Q1.8,COVR$Q1.11,COVR$Q1.14,COVR$Q1.21,COVR$Q1.22)) + 2
PCS2 = PCS
PCS[is.na(PCS)] = 0
PCS = rowSums(PCS)
PCS[rowSums(is.na(PCS2)) == 5] = NA
PCS[is.na(PCS)] = mean(PCS)
rm(PCS2)

#threats at admission (0 = no, 1 = yes)
Threats = cbind(COVR$QREAS.20,COVR$QREAS.21) - 1
T = Threats
Threats[is.na(Threats)] = 0
Threats = rowSums(Threats)
Threats[Threats > 1] = 1
Threats[rowSums(is.na(T)) == 2] = NA
Threats[is.na(Threats)] = varMode(Threats)
rm(T)

#number of negative relationships [0,8]
NegRel = cbind(COVR$Q10.10N,COVR$Q10.11N,COVR$Q10.12N,COVR$Q10.13N)
NegRel = rowMeans(NegRel)
#no missing data
```

C.2.2 Statistical Analyses

Below is the code for constructing the main effects logistic regression model, the discriminant function models, and the classification trees in MatLab. The dataset containing the predictor variables is called `COVRData`; the response variable, `Violence`.

Cross Validation

The code used to create the training and testing samples for cross-validation.

```
cvpart = cvpartition(Violence, 'holdout', .3);
XTrain = COVRData(training(cvpart),:);
YTrain = Violence(training(cvpart),:);
XTest = COVRData(test(cvpart),:);
YTest = Violence(test(cvpart),:);
sum(YTrain)/size(YTrain,1) %BRtrain
sum(YTest)/size(YTest,1) %BRtest
```

MELR Model

```
MELR = fitglm(COVRData(:,1:18), Violence, 'distr', 'binomial');
%95% CI
coefCI(MELR)
%odds ratio
OR = exp(MELR.Coefficients.Estimate);
UBor = exp(UB);
LBor = exp(LB);
%estimated probabilities
probs = MELR.Fitted.Probability
%ROC analysis
[fpr, tpr, ~, auc] = perfcurve(Violence, probs, 1);
```

C.2.3 Discriminant Analysis

```
%linear discriminant analysis
LDAmodel_lin = ClassificationDiscriminant.fit(COVRData, Violence);
%error measures
resubLoss(LDAmodel_lin)
cvmodel = crossval(LDAmodel_lin, 'leaveout', 'on');
kfoldLoss(cvmodel)
%predictions
```

```
[label, score, cost] = predict(LDAmodel_lin, COVRData);
%ROC analysis
[fpr, tpr, ~, auc] = perfcurve(Violence, score(:,2), 1);
%quadratic classifier
LDAmodel_quad = ClassificationDiscriminant.fit(COVRData, Violence, ...
    'DiscrimType', 'quadratic');
```

C.2.4 Decision Trees

```
%defining categorical predictors
CatVars = [6:7, 9:10, 13:25, 30];
%single tree
ctree = ClassificationTree.fit(COVRData, Violence, 'MinLeaf', 42, ...
    'CategoricalPredictors', CatVars, 'PredictorNames', VarNames);
%TreeBagger model
%fitting 1000 trees
ctreeBag = TreeBagger(1000, XTrain, YTrain, 'MinLeaf', 10, ...
    'CategoricalPredictors', CatVars, 'nprint', 100);
%OOB (Out-Of-Bag; observations not used in contruction of model)
ctreeBagOOB = TreeBagger(1000, COVRData, Violence, 'OOBVarImp', 'on', ...
    'MinLeaf', 10, 'CategoricalPredictors', CatVars, 'nprint', 100');
%determining importance of variables using OOB permuations
[sortedPermutedVarDeltaError, sortedVars] = ...
    sort(ctreeBagOOB.OOBPermutedVarDeltaError, 'descend');
%keeping most important variables based on OOB errors
topVars = sortedVars(sortedPermutedVarDeltaError > .01);
%final model
ctreeBagFinal = TreeBagger(10000, XTrain(:,topVars), YTrain, 'MinLeaf', 10, ...
    'OOBPred', 'on', 'CategoricalPredictors', CatVars(topVars), ...
    'nprint', 1000, 'PredictorNames', VarNames(topVars), ...
%applying costs
BR = sum(Violence)/length(Violence); %base rate
twoBR = (1-2*BR)/(2*BR);
costMatrix = [0 1; twoBR, 0]; costs
%e.g.,
ctreeBagCosts = TreeBagger(1000, XTrain, YTrain, 'MinLeaf', 10, ...
    'CategoricalPredictors', CatVars, 'nprint', 100, ...
    'cost', costMatrix);
%ROC analysis
[fpr, tpr, ~, auc] = perfcurve(YTest, YTestPredScores(:,2), 1);
%estimated cost function
```

```
[fpr, ecost, t] = perfcurve(YTest, YTestPredScores(:,2), 1, 'ycrit', ...
    'ecost', 'cost', [0, 1; 1, 0]);
```

SPSS Syntax

The following syntax was used in SPSS Version 21.0 to construct a CHAID decision tree model.

```
* Decision Tree.
TREE Violence [n] BY BISnp [s] BPRSa [n] BPRSh [n] BPRSt [s] ChildAbuse [s] Emp [n] DadDrug [n]
    PriorArr [s] GranDel [n] LegalStatus [n] SNMHP [s] NASb [s] Consc [n] PCL [n] DrugAbuse [n]
    tco [n] FantEsc [n] FantSing [n] RecViol2 [n] SubAbuse [n] Suicide [n] DadArr [n]
    HeadInj [n] FantTarg [n] Schiz [n] Age [s] Function [n] PropCrime [n] PCS [n] Threats [n]
    NegRel [s]
  /TREE DISPLAY = TOPDOWN NODES = STATISTICS BRANCHSTATISTICS = YES NODEDEFS = YES SCALE = AUTO
  /DEPCATEGORIES USEVALUES = [VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /METHOD TYPE = CHAID
  /GROWTHLIMIT MAXDEPTH = 100 MINPARENTSIZE = 100 MINCHILDSIZE = 50
  /VALIDATION TYPE = CROSSVALIDATION(10) OUTPUT = BOTHSAMPLES
  /CHAID ALPHASPLIT = 0.05 ALPHAMERGE = 0.05 SPLITMERGED = NO CHISQUARE = PEARSON CONVERGE = 0.001
    MAXITERATIONS = 100 ADJUST = BONFERRONI INTERVALS = 10
  /COSTS EQUAL
  /MISSING NOMINALMISSING = MISSING.
```

To add misclassification costs to the model, the following line replaced the penultimate line from the above syntax: `/COSTS CUSTOM= 1 1 [0] 1 2 [1]  2 1 [1.67] 2 2 [0]`

C.3 Brier Score Decomposition

From Chapter 6.

The `R` function below provides the Brier score and its decomposed parts as the output when provided a two-category dataset with the first column being the assigned probabilities and the second and third columns being the frequencies in the first and second category, respectively.

```
BrierScore <- function(forecastData) {
     #Brier Score decomposition for 2 category dataset
     #data should be in matrix with form
     #[predprob Y N]
     nr = rowSums(forecastData[,2:3])
     n = sum(nr)
     rep10 <- function(times)
          rep(c(1,0), times = times)
     y = unlist(apply(forecastData[,2:3], 1, rep10))
     yhat = rep(forecastData[,1], times = nr)
     pbar = sum(forecastData[,2])/n

     #Brier Score
     BS = sum((y - yhat)^2)/n
     #BS Decomposition
     #reliability
     Rel = sum(nr*(forecastData[,1] - forecastData[,2]/nr)^2)/n
     #resolution
     Res = sum(nr*(forecastData[,2]/nr - pbar)^2)/n
     #uncertainty
     Unc = pbar*(1-pbar)
     #uncertainty is minimal Brier score using naive prediction (i.e., using base rates)

     #check
     ifelse(round(BS - (Rel - Res + Unc), 10) == 0,
            return(list = c(BrierScore = BS, Reliability = Rel, Resolution = Res,
            Uncertainty = Unc)), "Brier score decomposition failed")
}
```

C.4 Kelley True Score Estimation

From Chapter 6.

The R code below simulates sensitivity values for eleven hypothetical prediction methods. The reliability is estimated using Hoyt's method (Hoyt, 1941) and Kelley's True Score Estimator (Kelley, 1923) is used to estimate the sensitivities and the mean squared error is compared to estimation using the observed values. For this example, the number of obser-

vations is set at 20.

```
require(MASS)
#hypothetical sensitivities of violence prediction methods
sens = matrix(seq(60, 80, 2)/100)
rsuccess <- function(p) sample(1:0, 20, replace = TRUE, prob = c(p, 1 - p))
set.seed(917)
data = apply(sens, 1, rsuccess)

#Hoyt, C. 1941. Test reliability estimated by analysis of variance.
n = nrow(data); k = ncol(data)
StudMean = matrix(rep(colMeans(data), each = n), n, k)
ItemMean = matrix(rep(rowMeans(data)), n, k)
OverMean = matrix(mean(data), n, k)
Y = data - ItemMean - StudMean + OverMean
S0 = sum(Y^2)
S1 = sum((data - ItemMean)^2)
S2 = S1 - S0
(r = ((n-1)*S2 - S0)/((n-1)*S2))

#Kelley True Score Estimation
X = StudMean[1,]
T = r*X + (1 - r)*mean(X)

#Mean Squared Error (Kelley True Score Estimator)
mean((T - sens)^2)
#Observed Score Estimator
mean((X - sens)^2)
```

C.5 Ridge Logistic Regression

The R code below fits the VRAS data using a logistic regression model fit using maximum likelihood estimation with and without a shrinkage penalty (i.e., ridge regression estimator). The variable names refer to those created in the script above (under Preprocessing the Data).

```
LogRegData =  cbind(Violence, BISnp, BPRSa, BPRSh, BPRSt,
                    ChildAbuse, Emp, DadDrug, PriorArr, GranDel, LegalStatus, SNMHP, NASb,
```

```
                    Consc, PCL, DrugAbuse, tco, FantEsc, FantSing)

LogRegData = as.data.frame(scale(LogRegData, center = F))
colnames(LogRegData) = c('Violence', 'BISnp', 'BPRSa', 'BPRSh', 'BPRSt', 'ChildAbuse',
                         'Emp', 'DadDrug', 'PriorArr', 'GranDel', 'LegalStatus', 'SNMHP',
                         'NASb', 'Consc', 'PCL', 'DrugAbuse', 'tco', 'FantEsc', 'FantSing')

#logistic regression model without shrinkage penalty
MainEffectsLogisticModel = glm(as.factor(Violence) ~ ., data = LogRegData,
                               family = binomial(logit))

#10 fold CV for GLM
require(boot)
#estimated probability of violence greater than 50%
MEMcost.5 = function(r, pi=0) mean(abs(r-pi) > 0.5)
cvMELR10 = cv.glm(LogRegData, MainEffectsLogisticModel, MEMcost.5, K = 10)
cvMELR10$delta[1] #K=10-fold error
RLMcvMER$cvm[which.min(RLMcvMER$cvm)] #K=10-fold error

#logistic regression model with shrinkage penalty
require(glmnet)
RidgeLogisticModel = glmnet(as.matrix(LogRegData[,2:19]), as.factor(Violence),
                            family = 'binomial', alpha = 0, standardize = F)
#10 fold CV for ridge model
#misclassification error
RLMcvMER = cv.glmnet(as.matrix(LogRegData[,2:19]), as.factor(Violence), family = 'binomial',
                         type.measure = 'class', alpha = 0, standardize = F)
#min cv error
RLMcvMER$lambda.min

#model using CV lambda
RidgeLogisticModel = glmnet(as.matrix(LogRegData[,2:19]), as.factor(Violence),
                            family = 'binomial', alpha = 0, standardize = F,
                            lambda = RLMcvMER$lambda.min)
```

References

Abbott, B. R. (2011). Throwing the baby out with the bath water: Is it time for clinical judgment to supplement actuarial risk assessment? *Journal of the American Academy of Psychiatry and the Law*, *39*, 222–230.

Addington v. Texas, 441 U.S. 418. (1979).

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., ... Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*, 341–382.

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*, 119–126.

Aharoni, E., Vincent, G. M., Harenski, C. L., Calhoun, V. D., Sinnott-Armstrong, W. P., Gazzaniga, M. S., & Kiehl, K. A. (2013). Neuroprediction of future rearrest. *Proceedings of the National Academy of Sciences*, *110*, 6223–6228.

Allan, M., Grace, R. C., Rutherford, B., & Hudson, S. M. (2007). Psychometric assessment of dynamic risk factors for child molesters. *Sexual Abuse: A Journal of Research and Treatment*, *19*, 347–367.

American Psychiatric Association. (1983). *Barefoot v. Estelle: Amicus curiae brief.*

American Psychiatric Publishing. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Arlington, VA: American Psychiatric Publishing.

Andrews, D. A. (1988). *The Level of Supervision Inventory (LSI)*. Toronto, Canada: Ontario Ministry of Correctional Services.

Andrews, D. A., & Bonta, J. (1995). *LSI-R: The Level of Service Inventory–Revised: Manual*. Toronto, Canada: Multi-Health Systems, Inc.

Andrews, D. A., & Bonta, J. (1998). Level of Service Inventory–Screening Version. *Multi-Health Systems, Inc.*.

Anonymous. (2012, December). *"I am Adam Lanza's psychiatrist": A response from the mental health trenches to "I am Adam Lanza's mother"*. Retrieved from `http://www.xojane.comt`

Augimeri, L. K., Koegl, C. J., Webster, C. D., & Levene, K. S. (2001). *Early assessment risk list for boys: EARL-20B, Version 2*. Toronto, Canada: Earlscourt Child and Family Centre.

Augimeri, L. K., Webster, C. D., Koegl, C. J., & Levene, K. S. (1998). *Early assessment risk list for boys: EARL-20B, Version 1 (Consultation edition)*. Toronto, Canada: Earlscourt Child and Family Centre.

Babchishin, K. M., Hanson, R. K., & Helmus, L. (2011). *The RRASOR, Static-99R and Static-2002R all add incrementally to the prediction of recidivism among sex offenders.* Ottawa, Canada: Public Safety Canada.

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*, 411–421.

Banks, S., Robbins, P. C., Silver, E., Vesselinov, R., Steadman, H. J., Monahan, J., ... Roth, L. H. (2004). A multiple-models approach to violence risk assessment among people with mental disorder. *Criminal Justice and Behavior*, *31*, 324–340.

Barbaree, H. E., Seto, M. C., Langton, C. M., & Peacock, E. J. (2001). Evaluating the predictive accuracy of six risk assessment instruments for adult sex offenders. *Criminal Justice and Behavior*, *28*, 490–521.

Barefoot v. Estelle, 463 U.S. 880. (1983).

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*, 211–233.

Bartel, P. A., Forth, A. E., & Borum, R. (2003). *Development and concurrent validation of the Structured Assessment of Violence Risk in Youth (SAVRY)* Unpublished manuscript.

Bartosh, D. L., Garby, T., Lewis, D., & Gray, S. (2003). Differences in the predictive validity of actuarial risk assessments in relation to sex offender type. *International Journal of Offender Therapy and Comparative Criminology*, *47*, 422–438.

Bauer, W. (1970). The other side of the coin. *Illinois Medical Journal*, *137*, 158–161.

Baxstrom v. Herold, 383 U.S. 107. (1966).

Bechman, D. C. (2001). Sex offender civil commitments: Scientists or physics. *Criminal Justice*, *16*, 24–33.

Begg, C. B. (1987). Biases in the assessment of diagnostic tests. *Statistics in Medicine*, *6*, 411–423.

Beis, E. (1983). State involuntary commitment statutes. *Mental Disability Law Reporter*, *7*, 358–369.

Belson, K. (2013, December 25). *New tests for brain trauma create hope, and skepticism.* Retrieved from `http://www.nytimes.com`

Berger, L. S., & Dietrich, S. G. (1979). Clinical prediction of dangerousness: Logic of the process. *International Journal of Offender Therapy and Comparative Criminology*, *23*, 35–46.

Berk, R. (2009). The role of race in forecasts of violent crime. *Race and Social Problems*, *1*, 231–242.

Berk, R. (2011). Asymmetric loss functions for forecasting in criminal justice settings. *Journal of Quantitative Criminology*, *27*, 107–123.

Berk, R. (2012). *Criminal justice forecasts of risk: A machine learning approach.* New York, NY: Springer.

Berkson, J. (1946). Limitations of the application of fourfold tables to hospital data. *Biometrics Bulletin*, *2*, 47–53.

Berkson, J. (1947). "Cost-utility" as a measure of the efficiency of a test. *Journal of the American Statistical Association*, *42*, 246–255.

Berlin, F. S., Galbreath, N. W., Geary, B., & McGlone, G. (2003). The use of actuarials at civil commitment hearings to predict the likelihood of future sexual violence. *Sexual Abuse: A Journal of Research and Treatment*, *15*, 377–382.

Biel, L. (1995, July). *Groups expel Texas psychiatrist known for murder cases.* Dallas Morning News.

Black, B. J., & Glick, S. J. (1952). *Recidivism at the Hawthorne-Cedar Knolls school: Predicted vs. actual outcome for delinquent boys.* New York, NY: Jewish Board of Guardians.

Blackstone, W. (1794). *Commentaries on the Laws of England in one volume* (W. H. Browne, Ed.).

Blair, P. R., Marcus, D. K., & Boccaccini, M. T. (2008). Is there an allegiance effect for assessment instruments? actuarial risk assessment as an exemplar. *Clinical Psychology: Science and Practice*, *15*, 346–360.

Bloom, J. D., & Rogers, J. L. (1987). The legal basis of forensic psychiatry: Statutorily mandated psychiatric diagnoses. *American Journal of Psychiatry*, *144*, 847–853.

Bloom, P. (2013, July 21). *Natural born killers: 'The anatomy of violence' by Adrian Raine.* Retrieved from `http://www.nytimes.com`

Boccaccini, M. T., Murrie, D. C., Mercado, C., Quesada, S., Hawes, S., Rice, A. K., & Jeglic, E. L. (2012). Implications of Static-99 field reliability findings for score use and reporting. *Criminal Justice and Behavior*, *39*, 42–58.

Boccaccini, M. T., Turner, D. B., & Murrie, D. C. (2008). Do some evaluators report consistently higher or lower PCL-r scores than others? findings from a statewide sample of sexually violent predator evaluations. *Psychology, Public Policy, and Law*, *14*, 262.

Boer, D. P., Hart, S. D., Kropp, P. R., & Webster, C. D. (1997). *Manual for the Sexual Violence Risk-20: Professional guidelines for assessing risk of sexual violence.* Burnaby, Canada: Simon Fraser University.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*, 214–234.

Bonta, J. (1996). Choosing correctional options that work: Defining the demand and evaluating the supply. In A. Harland (Ed.), (pp. 18–32). Thousand Oaks, CA: Sage Publications, Inc.

Bonta, J., Law, M., & Hanson, K. (1998). The prediction of criminal and violent recidivism among mentally disordered offenders: A meta-analysis. *Psychological Bulletin*, *123*, 123–142.

Borum, R., Bartel, P. A., & Forth, A. E. (2005). Mental health screening and assessment in juvenile justice. In T. Grisso, G. Vincent, & D. Seagrave (Eds.), (pp. 311–323). New York, NY: Guiliford Press.

Borum, R., Bartel, P. A., & Forth, A. E. (2006). *Manual for the Structured Assessment of Violence Risk in Youth (SAVRY).* Odessa, NY: Psychological Assessment Resources.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., ... Lijmer, J. G. (2003). The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Clinical Chemistry*, *49*, 7–18.

Botelho, G., & Sterling, J. (2013, September 26). *FBI: Navy Yard shooter 'delusional' said 'low frequency attacks' drove him to kill.* Retrieved from http://www.cnn.com

Bratu, B. (2012, December). *Connecticut school shooting is second worst in US history.* Retrieved from http://usnews.nbcnews.com

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *26*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth & Brooks.

Brenner, H., & Gefeller, O. (1997). Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Statistics in Medicine*, *16*, 981–991.

Breslow, J. M. (2013, February). *For Adam Lanza, a debated diagnosis that meant "more to be worried about".* Retrieved from http://www.pbs.org

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.

Bryan, A. (2012, December). *TIMELINE: Connecticut elementary school shooting updates.* Retrieved from http://wtvr.com

Buchanan, A. (1999). Risk and dangerousness. *Psychological Medicine*, *29*, 465–473.

Buck v. Thaler, 132 S. Ct. 32. (2011).

Buffington-Vollum, J., Edens, J. F., Johnson, D. W., & Johnson, J. K. (2002). Psychopathy as a predictor of institutional misbehavior among sex offenders: A prospective replication. *Criminal Justice and Behavior*, *29*, 497–511.

Bump, P. (2013a, September 17). *Where will be another mass shooting. This is what the data tells us about it.* Retrieved from http://www.theatlanticwire.com

Bump, P. (2013b, September 18). *Spokane shouldn't be the only place scared of the next mass shooting.* Retrieved from http://www.theatlanticwire.com

Burgess, E. W. (1928). Workings of the indeterminate sentence law and parole system in Illinois: A report to the Honorable Hinton G. Clabaugh. In A. A. Bruce (Ed.), (pp. 246–249). Springfield, IL: Illinois State Board of Parole.

Butz-Whittaker, J. W., Strassberg, D. S., & the Center for Family Development. (2001). *The Sex Offender Treatment Outcome PredictorStatic/Dynamic (SOTOP-S/D).* Poster presented at the 20th Annual Conference of the Association for the Treatment of Sexual Abusers, San Antonio, TX.

Campbell, M. A., French, S., & Gendreau, P. (2007). *Assessing the utility of risk assessment tools and personality measures in the prediction of violent recidivism for adult offenders.* Ottawa, Canada: Department of Public Safety and Emergency Preparedness.

Candiotti, S., & Aarthun, S. (2012, December). *Police: 20 children among 26 victims of Connecticut school shooting.* Retrieved from http://edition.cnn.com

Cantor, P. D., & Sherman, P. R. (1965). Hospitalization of the Mentally Ill in the District of Columbia. *American University Law Review*, *15*, 203–222.

Capwell, D. F. (1945). Personality patterns of adolescent girls: II. Delinquents and non-delinquents. *Journal of Applied Psychology*, *29*, 289–297.

Carroll, J. S., Wiener, R. L., Coates, D., Galegher, J., & Alibrio, J. J. (1982). Evaluation, diagnosis, and prediction in parole decision making. *Law and Society Review*, 199–228.

Carroll, L. (1875). *Through the looking glass, and what Alice found there.* Macmillan & Company.

Cassels, A. (2012). *Seeking sickness: Medical screening and the misguided hunt for disease.* Vancouver, Canada: Greystone Books.

Cassels, A., Van Wiltenburg, J., & Armstrong, W. L. (2009). *What's in a scan?: How well are consumers informed about the benefits and harms related to screening technology (CT and PET scans) in Canada.* Ottawa, Canada: Canadian Centre for Policy Alternatives.

Catchpole, R. E., & Gretton, H. M. (2003). The predictive validity of risk assessment with violent young offenders: A 1-year examination of criminal outcome. *Criminal Justice and Behavior*, *30*, 688–708.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, *72*.

Clark, W. W. (1920). Success record of delinquent boys in relation to intelligence. *Journal of Delinquency*, *5*, 174–182.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*, 404–413.

Cloud, J. (2011, January 15). *The troubled life of Jared Loughner.* Retrieved from `http://content.time.com`

Coble v. State, 330 S.W. 3d 253. (2010).

Cocozza, J. J., Melick, M. E., & Steadman, H. J. (1978). Trends in violent crime among ex-mental patients. *Criminology*, *16*, 317–334.

Cocozza, J. J., & Steadman, H. J. (1974). Some refinements in the measurement and prediction of dangerous behavior. *American Journal of Psychiatry*, *131*, 1012–1014.

Cocozza, J. J., & Steadman, H. J. (1975). The failure of psychiatric predictions of dangerousness: Clear and convincing evidence. *Rutgers Law Review*, *29*, 1084–1101.

Cocozza, J. J., & Steadman, H. J. (1978). Prediction in psychiatry: An example of misplaced confidence in experts. *Social Problems*, 265–276.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York, NY: Academic Press.

Cohen, M. L., Groth, A. N., & Siegel, R. (1978). The clinical prediction of dangerousness. *Crime and Delinquency*, *24*, 28–39.

Coid, J., Yang, M., Ullrich, S., Zhang, T., Roberts, A., Roberts, C., ... Farrington, D. (2007). *Predicting and understanding risk of re-offending: The Prisoner Cohort Study.*

Conner, B. T., Hellemann, G. S., Ritchie, T. L., & Noble, E. P. (2010). Genetic, personality, and environmental predictors of drug use in adolescents. *Journal of Substance Abuse Treatment*, *38*, 178–190.

Consortium for risk-based firearm policy. (2013). *Guns, public health, and mental illness: An Evidence-Based Approach for State Policy.*

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, *115*, 928–935.

Cook, N. R. (2008). Statistical evaluation of prognostic versus diagnostic models: Beyond the roc curve. *Clinical Chemistry*, *54*, 17–23.

Cooke, D. J., & Michie, C. (2010). Limitations of diagnostic precision and predictive utility in the individual case: A challenge for forensic practice. *Law and Human Behavior*, *34*, 259–274.

Copas, J., & Marshall, P. (1998). The Offender Group Reconviction Scale: A statistical reconviction score for use by probation officers. *Journal of the Royal Statistical Society*, *47*, 159–171.

Corrado, M. L. (1996). Punishment and the wild beast of prey: The problem of preventive detention. *The Journal of Criminal Law and Criminology*, *86*, 778–814.

Cross v. Harris, 4186 F. 2d 1095. (1969).

Cunningham, M. D., & Reidy, T. J. (1999). Don't confuse me with the facts: Common errors in violence risk assessment at capital sentencing. *Criminal Justice and Behavior*, *26*, 20–43.

Cunningham, M. D., & Sorensen, J. R. (2006). Actuarial models for assessing prison violence risk revisions and extensions of the Risk Assessment Scale for Prison (RASP). *Assessment*, *13*, 253–265.

Cunningham, M. D., & Sorensen, J. R. (2007). Predictive factors for violent misconduct in close custody. *The Prison Journal*, *87*, 241–253.

Dash v. Mitchell, 365 F. Supp. 1292. (1965).

Daubert v. Merrel Dow Pharmaceuticals, Inc., 509 U.S. 579. (1993).

Davis-Stober, C. P., Dana, J., & Budescu, D. V. (2010). A constrained linear estimator for multiple regression. *Psychometrika*, *75*, 521–541.

Dawes, R. M. (1962). A note on base rates and psychometric efficiency. *Journal of Consulting Psychology*, *26*, 422–424.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*, 571–582.

Dawes, R. M. (1986). Forecasting one's own preference. *International Journal of Forecasting*, *2*, 5–14.

Dawes, R. M. (1993). Prediction of the future versus an understanding of the past: A basic asymmetry. *The American Journal of Psychology*, *106*, 1–24.

Dawes, R. M. (2002). The ethics of using or not using statistical prediction rules in psychological practice and related consulting activities. *Philosophy of Science*, *69*, S178–S184.

Dawes, R. M. (2005). The ethical implications of Paul Meehl's work on comparing clinical versus actuarial prediction methods. *Journal of Clinical Psychology*, *61*, 1245–1255.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*, 95–106.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1674.

de Vogel, V., & de Ruiter, C. (2005). The HCR-20 in personality disordered female offenders: A comparison with a matched sample of males. *Clinical Psychology and Psychotherapy*, *12*, 226–240.

de Vogel, V., & de Ruiter, C. (2006). Structured professional judgment of violence risk in forensic clinical practice: A prospective study into the predictive validity of the Dutch HCR-20. *Psychology, Crime and Law*, *12*, 321–336.

de Vogel, V., de Ruiter, C., van Beek, D., & Mead, G. (2004). Predictive validity of the SVR-20 and Static-99 in a Dutch sample of treated sex offenders. *Law and Human Behavior*, *28*, 235–251.

Deming, A. (2008). Sex offender civil commitment programs: Current practices, characteristics, and resident demographics. *Journal of Psychiatry & Law*, *36*, 439–461.

Dershowitz, A. M. (1967). Psychiatry in the legal process: A knife that cuts both ways. *Judicature*, *51*, 370–377.

Dershowitz, A. M. (1970). The law of dangerousness: Some fictions about predictions. *Journal of Legal Education*, *23*, 24–56.

Dershowitz, A. M. (1971). Imprisonment by judicial hunch. *American Bar Association Journal*, *57*, 560–564.

Diamond, B. L. (1974). The psychiatric prediction of dangerousness. *University of Pennsylvania Law Review*, *123*, 439–452.

Dilulio, J. J., Jr. (1995, November 27). The coming of the super-predator. *The Weekly Standard*.

District of Columbia Court Reform and Criminal Procedure Act. (1970).

District of Columbia Hospitalization of the Mentally Ill Act, 79 Stat. 944. (1965).

Dix, G. F. (1976). Civil commitment of the mentally ill and the need for data on the prediction of dangerousness. *American Behavioral Scientist*, *19*, 318–334.

Doe v. Poritz, 142 N.J. 1. (1995).

Dolan, M., & Doyle, M. (2000). Violence risk prediction: Clinical and actuarial measures and the role of the Psychopathy Checklist. *The British Journal of Psychiatry*, *177*, 303–311.

Donaldson v. O'Conner, 493 F. 2d 507. (1975).

Doren, D. M. (1998). Recidivism base rates, predictions of sex offender recidivism, and the "sexual predator" commitment laws. *Behavioral Sciences and the Law*, *16*, 97–114.

Doren, D. M. (2000). Evidentiary issues, actuarial scales, and sex offender civil commitments. *Sex Offender Law Report*, *1*, 65–66, 78–79.

Doren, D. M. (2004a). Stability of the interpretative risk percentages for the RRASOR and Static-99. *Sexual Abuse: A Journal of Research and Treatment*, *16*, 25–36.

Doren, D. M. (2004b). Toward a multidimensional model for sexual recidivism risk. *Journal of Interpersonal Violence*, *19*, 835–856.

Douglas, K. S., Guy, L. S., & Hart, S. D. (2009). Psychosis as a risk factor for violence to others: A meta-analysis. *Psychological Bulletin*, *135*, 679–706.

Douglas, K. S., Ogloff, J. R. P., & Hart, S. D. (2003). Evaluation of a model of violence risk assessment among forensic psychiatric patients. *Psychiatric Services*, *54*, 1372–1379.

Douglas, K. S., & Skeem, J. L. (2005). Violence risk assessment: Getting specific about being dynamic. *Psychology, Public Policy, and Law*, *11*, 347–383.

Douglas, K. S., Yeomans, M., & Boer, D. P. (2005). Comparative validity analysis of multiple measures of violence risk in a sample of criminal offenders. *Criminal Justice and Behavior*, *32*, 479–510.

Doyle, M., Dolan, M., & McGovern, J. (2002). The validity of North American risk assessment tools in predicting in-patient violent behaviour in England. *Legal and Criminological Psychology*, *7*, 141–154.

Doyle, M., & Logan, C. (2012). Operationalizing the assessment and management of violence risk in the short-term. *Behavioral Sciences and the Law*, *30*, 406–419.

Doyle, M., Shaw, J., Carter, S., & Dolan, M. (2010). Investigating the validity of the Classification of Violence Risk in a UK sample. *International Journal of Forensic Mental Health*, *9*, 316–323.

Duncan, O. D., Ohlin, L. E., Reiss, A. J., Jr., & Stanton, H. R. (1953). Formal devices for making selection decisions. *American Journal of Sociology*, 573–584.

Duwe, G., & Freske, P. J. (2012). Using logistic regression modeling to predict sexual recidivism: The Minnesota Sex Offender Screening Tool-3 (MnSOST-3). *Sexual Abuse: A Journal of Research and Treatment*, *24*, 350–377.

Edens, J. F., Skeem, J. L., & Douglas, K. S. (2006). Incremental validity analyses of the Violence Risk Appraisal Guide and the Psychopathy Checklist: Screening version in a civil psychiatric sample. *Assessment*, *13*, 368–374.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*, 1–26.

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, *68*, 117–130.

Elbogen, E. B., & Johnson, S. C. (2009). The intricate link between violence and mental disorder: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. *Archives of general Psychiatry*, *66*, 152–161.

Enebrink, P., Långström, N., & Gumpert, C. H. (2006). Predicting aggressive and disruptive behavior in referred 6-to 12-year-old boys prospective validation of the EARL-20B risk/needs checklist. *Assessment*, *13*, 356–367.

Ennis, B. J., & Litwack, T. R. (1974). Psychiatry and the presumption of expertise: Flipping coins in the courtroom. *California Law Review*, *62*, 693–752.

Epperson, D. L., Kaul, J. D., & Hesselton, D. (1998). *Final report on the development of the Minnesota Sex Offender Screening Tool–Revised (MnSOST-R).* Paper presentated at the 17th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, Vancouver, Canada.

Epperson, D. L., Kaul, J. D., & Huot, S. J. (1995). *Predicting risk of recidivism for incarcerated sex offenders: Updated development on the Sex Offender Screening Tool (SOST).* Paper presented at the 14th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, New Orleans, LA.

Espada v. State, No. AP-75.219. (2008).

Estelle v. Smith, 451 US 454. (1981).

Estrada v. State, 313 S.W. 3d. 274. (2010).

Ewing, C. P. (1985). Schall v. Martin: Preventive detention and dangerousness through the looking glass. *Buffalo Law Review*, *34*, 173–226.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the Subjective Numeracy Scale. *Medical Decision Making*, *27*, 672–680.

Faigman, D. L., Blementhal, J. A., Cheng, E. K., Mnookin, J. L., Murphy, E. E., & Sanders, J. (Eds.). (2013). *Modern scientific evidence: The law and science of expert testimony* (2013–2014 ed., Vol. 2: Social & Behavioral Sciences). Eagan, MN: Thomson Reuters.

Faust, D., & Nurcombe, B. (1989). Improving the accuracy of clinical judgment. *Psychiatry: Interpersonal and Biological Processes*, 197–208.

Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24,827 people: Systematic review and meta-analysis. *British Medical Journal*, *345*, e4692.

Federal Rules of Evidence. (2013).

Fergusson, D. M., Fifield, J. K., & Slater, S. W. (1977). Signal detectability theory and the evaluation of prediction tables. *Journal of Research in Crime and Delinquency*, *14*, 237–246.

Ferri, C., Hernández-Orallo, J., & Flach, P. A. (2011). Brier curves: A new cost-based visualisation of classifier performance. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 585–592).

Fienberg, S. E. (1989). *The evolving role of statistical assessments as evidence in the courts.* New York: Springer-Verlag.

Fienberg, S. E., & Kadane, J. B. (1983). The presentation of Bayesian statistical analyses in legal proceedings. *The Statistician*, 88–98.

Firestone, P., Bradford, J. M., McCoy, M., Greenberg, D. M., Curry, S., & Larose, M. R. (2000). Prediction of recidivism in extrafamilial child molesters based on court-related assessments. *Sexual Abuse: A Journal of Research and Treatment*, *12*, 203–221.

Fitch, W. L. (2003). Sexually coercive behavior. In R. A. Prentky, E. S. Janus, & M. C. Seto (Eds.), (pp. 489–501). New York, NY: The New York Academy of Sciences.

Flores v. Johnson, 210 F. 3d 456. (2000).

Floyd v. The City of New Yor, No. 08 Civ. 1034. (2013a).

Floyd v. The City of New Yor, Nos. 08 Civ. 1034, 12 Civ. 2274. (2013b).

Follman, M. (2012a, November 9). *Mass shootings: Maybe what we need is a better mental-health policy.* Retrieved from `http://www.motherjones.com`

Follman, M. (2012b, December 15). *More guns, more mass shootings—coincidence?* Retrieved from `http://www.motherjones.com`

Follman, M., Aronsen, G., Pan, D., & Caldwell, M. (2012, December 28). *US mass shootings, 1982-2012: Data from Mother Jones' investigation.* Retrieved from `http://www.motherjones.com`

Foote, C. (1970a). Comments on preventive detention. *Journal of Legal Education*, *23*, 48–55.

Foote, C. (1970b). Preventive detention—what is the issue? *The Prison Journal*, *50*, 3–11.

Forth, A. E., Kosson, D. S., & Hare, R. D. (2003). *Hare Psychopathy Checklist: Youth Version (PCL:YV).* Toronto, Canada: Multi-Health Systems, Inc.

Foucha v. Louisiana, 504 U.S. 71. (1992).

Fox, J. A. (2012, August 6). *Are mass shootings becoming more common in the United States?* Retrieved from `http://boston.com`

Frederick, R. I., & Bowden, S. C. (2009). The test validation summary. *Assessment*, *16*, 215–236.

Frye v. United States, 293 F. 1013. (D.C. Cir. 1923).

Furda v. State, 997 A. 2d 856. (2010).

Furnas, A. (2012, April 17). *Homeland security's 'pre-crime' screening will never work.* Retrieved from `http://www.theatlantic.com`

Furtado, C. D., Aguirre, D. A., Sirlin, C. B., Dang, D., Stamato, S. K., Lee, P., ... Casola, G. (2005). Whole-body CT screening: Spectrum of findings and recommendations in 1192 patients. *Radiology*, *237*, 385–394.

Gardner, W., Lidz, C. W., Mulvey, E. P., & Shaw, E. C. (1996a). Clinical versus actuarial predictions of violence of patients with mental illnesses. *Journal of Consulting and Clinical Psychology*, *64*.

Gardner, W., Lidz, C. W., Mulvey, E. P., & Shaw, E. C. (1996b). A comparison of actuarial methods for identifying repetitively violent patients with mental illnesses. *Law and Human Behavior*, *20*, 35–48.

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology*, *34*, 575–608.

General Electric Co. v. Joiner, 522 U.S. 136. (1997).

George, R. (2013, September 17). *Why are mass shootings increasing even while gun violence is decreasing?* Retrieved from `http://www.policymic.com`

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 53–96.

Gini, C. (1912). *Variabilità e mutabilità: Contributo allo studio delle distribuzioni e delle relazioni statistiche [Variability and mutability: Contribution to the study of distributions and report statistics].* Bologna, Italy: C. Cuppini.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*, 1012–1014.

Glaser, D. (1955). The efficacy of alternative approaches to parole prediction. *American Sociological Review*, *20*, 283–287.

Glueck, S., & Glueck, E. T. (1930). *500 criminal careers.* New York, NY: Alfred A Knopf.

Glueck, S., & Glueck, E. T. (1934). *One thousand juvenile delinquents.* Cambridge, MA: Harvard University Press.

Glueck, S., & Glueck, E. T. (1950). *Unraveling juvenile delinquency.* New York, NY: The Commonwealth Fund.

Glueck, S., & Glueck, E. T. (1968). *Delinquents and nondelinquents in perspective.* Cambridge, MA: Harvard University Press.

Goode, E. (2011, August 15). *Sending the police before there's a crime.* Retrieved from `http://www.nytimes.com`

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of American Statistical Association*, *49*, 732–762.

Gottfredson, S. D. (1987). Prediction: An overview of selected methodological issues. *Crime and Justice*, *9*, 21–51.

Gottfredson, S. D., & Moriarty, L. J. (2006). Statistical risk assessment: Old problems and new applications. *Crime and Delinquency*, *52*, 178–200.

Grady, D. (2013, September 18). *Signs may be evident in hindsight, but predicting violent behavior is tough.* Retrieved from http://www.nytimes.com

Grann, M., Belfrage, H., & Tengström, A. (2000). Actuarial assessment of risk for violence: Predictive validity of the VRAG and the historical part of the HCR-20. *Criminal Justice and Behavior*, *27*, 97–114.

Gray, N. S., Hill, C., McGleish, A., Timmons, D., MacCulloch, M. J., & Snowden, R. J. (2003). Prediction of violence and self-harm in mentally disordered offenders: A prospective study of the efficacy of HCR-20, PCL-R, and psychiatric symptomatology. *Journal of Consulting and Clinical Psychology*, *71*, 443–451.

Gray, S. (2012, December). *Sandy Hook gunman Adam Lanza shot his way through school door.* Retrieved from http://www.thetimes.co.uk

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York, NY: Wiley.

Greenberg, D. F. (1979). *Mathematical criminology.* New Brunswick, NJ: Rutgers University Press.

Gretton, H. M., McBride, M., Hare, R. D., O'Shaughnessy, R., & Kumka, G. (2001). Psychopathy and recidivism in adolescent sex offenders. *Criminal Justice and Behavior*, *28*, 427–449.

Grisso, T., & Appelbaum, P. S. (1992). Is it unethical to offer predictions of future violence? *Law and Human Behavior*, *16*, 621–633.

Grisso, T., Malamuth, N. M., Barbaree, H., Quinsey, V., & Knight, R. (2003). Sexually coercive behavior. In R. A. Prentky, E. S. Janus, & M. C. Seto (Eds.), (pp. 236–246). New York, NY: The New York Academy of Sciences.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, *2*, 293–323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological assessment*, *12*, 19–30.

Grubin, D. (1998). *Sex offending against children: Understanding the risk.* Police Research Series Paper 99. London, England.

Guggenmoos-Holzmann, I., & van Houwelingen, H. C. (2000). The (in)validity of sensitivity and specificity. *Statistics in Medicine*, *19*, 1783–1792.

Guy, L. S. (2008). *Performance indicators of the structured professional judgment approach for assessing risk for violence to others: A meta-analytic survey* (Doctoral dissertation, Simon Fraser University). Retrieved from http://summit.sfu.ca/item/9247

Hacker, F. J., & Frym, M. (1955). The sexual psychopath act in practice: A critical discussion. *California Law Review*, *43*, 766–780.

Hakeem, M. (1945). Prediction of criminality. *Federal Probation*, *9*, 31–38.

Hand, D. J., & Anagnostopoulos, C. (2013). When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern Recognition Letters*, *34*, 492–495.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36.

Hanson, R. K. (1997). *The development of a brief actuarial risk scale for sexual offense recidivism.* (User Report 97-04). Ottawa, Canada.

Hanson, R. K. (2002). Recidivism and age: Follow-up data from 4,673 sexual offenders. *Journal of Interpersonal Violence*, *17*, 1046–1062.

Hanson, R. K., & Howard, P. D. (2010). Individual confidence intervals do not inform decision-makers about the accuracy of risk assessment evaluations. *Law and Human Behavior*, *34*, 275–281.

Hanson, R. K., & Karl, R. (2003). *Notes on the development of Static-2002.* (Corrections Research User Report No. 2003-01). Ottawa, Canada.

Hanson, R. K., Lloyd, C. D., Helmus, L., & Thornton, D. (2012). Developing non-arbitrary metrics for risk communication: Percentile ranks for the Static-99/R and Static-2002/R sexual offender risk tools. *International Journal of Forensic Mental Health*, *11*, 9–23.

Hanson, R. K., & Morton-Bourgon, K. (2007). *The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis.* Corrections User Report No 2007-01. Public Safety and Emergency Preparedness, Ottawa, Canada..

Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior*, *24*, 119–136.

Harcourt, B. E. (2008). *Against prediction: Profiling, policing, and punishing in an actuarial age.* Chicago, IL: University of Chicago Press.

Hare, R. D. (1980). A research scale for the assessment of psychopathy in criminal populations. *Personality and Individual Differences*, *1*, 111–119.

Hare, R. D. (1991). *The Hare Psychopathy Checklist-Revised.* Toronto, Canada: Multi-Health Systems, Inc.

Hare, R. D. (1998). The Hare PCL-R: Some issues concerning its use and misuse. *Legal and Criminological Psychology*, *3*, 99–119.

Harris, G. (2003). Men in his category have a 50% likelihood, but which half is he in? Comments on Berlin, Galbreath, Geary, and McGlone. *Sexual Abuse: A Journal of Research and Treatment*, *15*, 389–392.

Harris, G. T., & Rice, M. E. (2003). Sexually coercive behavior. In R. A. Prentky, E. S. Janus, & M. C. Seto (Eds.), (pp. 198–210). New York, NY: The New York Academy of Sciences.

Harris, G. T., & Rice, M. E. (2007). Characterizing the value of actuarial violence risk assessments. *Criminal Justice and Behavior*, *34*, 1638–1658.

Harris, G. T., & Rice, M. E. (2013). Bayes and base rates: What is an informative prior for actuarial violence risk assessment? *Behavioral Sciences and the Law*, *31*, 103-124.

Harris, G. T., Rice, M. E., & Camilleri, J. A. (2004). Applying a forensic actuarial assessment (the Violence Risk Appraisal Guide) to nonforensic patients. *Journal of Interpersonal Violence*, *19*, 1063–1074.

Harris, G. T., Rice, M. E., & Cormier, C. A. (2002). Prospective replication of the Violence Risk Appraisal Guide in predicting violent recidivism among forensic patients. *Law and Human Behavior*, *26*, 377–394.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal Justice and Behavior*, *20*, 315–335.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (2008). Shall evidence-based risk assessment be abandoned? *British Journal of Psychiatry*, *192*, 154.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (2010). Allegiance or fidelity? A clarifying reply. *Clinical Psychology: Science and Practice*, *17*, 82–89.

Harris, G. T., Rice, M. E., Quinsey, V. L., Lalumiere, M. L., Boer, D. P., & Lang, C. (2003). A multisite comparison of actuarial risk instruments for sex offenders. *Psychological Assessment*, *15*, 413–425.

Hart, H. (1923). Predicting parole success. *Journal of the American Institute of Criminal Law and Criminology*, *14*, 405–413.

Hart, S. D. (2009). Evidence-based assessment of risk for sexual violence. *Chapman Journal of Criminal Justice*, *1*, 143–165.

Hart, S. D., & Cooke, D. J. (2013). Another look at the (im-)precision of individual risk estimates made using actuarial risk assessment instruments. *Behavioral Sciences and the Law*, 81–102.

Hart, S. D., Cox, D. N., & Hare, R. D. (1995). *The Hare Psychopathy Checklist: Screening Version* (1st ed.). Toronto, Canada: Multi-Health Systems, Inc.

Hart, S. D., Kropp, P. R., & Laws, D. R. (2003). *The risk for sexual violence protocol (rsvp): Structured professional guidelines for assessing risk of sexual violence.* Burnaby, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University and British Columbia Institute on Family Violence.

Hart, S. D., Michie, C., & Cooke, D. J. (2007). Precision of actuarial risk assessment instruments: Evaluating the 'margins of error' of group v. individual predictions of violence. *The British Journal of Psychiatry*, *190*, s60–s65.

Hart, S. D., Webster, C. D., & Menzies, R. J. (1993). A note on portraying the accuracy of violence predictions. *Law and Human Behavior*, *17*, 695–700.

Hartill, R., O'Cain, D., & Sutton, E. (2013, January). *Plot summary for Falling Down (1993)* [Review of *Falling Down*, produced by Timothy Harris, Arnold Kopelson, and Herschel Weingrod, directed by Joel Schumacher, 1993]. Retrieved from `http://www.imdb.com`

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Second ed.). New York, NY: Springer.

Hathaway, S. R., & Monachesi, E. D. (1951). The prediction of juvenile delinquency using the Minnesota Multiphasic Personality Inventory. *American Journal of Psychiatry*, *108*, 469–473.

Hathaway, S. R., & Monachesi, E. D. (1952). The Minnesota Multiphasic Personality Inventory in the study of juvenile delinquents. *American Sociological Review*, *17*, 704–710.

Healy, M. L., Gibney, J., Pentecost, C., Wheeler, M. J., & Sonksen, P. H. (in press). Endocrine profiles in 693 elite athletes in the postcompetition setting. *Clinical Endocrinology*.

Heilbrun, K., Douglas, K. S., & Yasuhara, K. (2009). Psychological science in the courtroom: Consensus and controversy. In J. L. Skeem, K. S. Douglas, & S. O. Lilienfeld (Eds.), (chap. Violence risk assessment: Core controversies). New York, NY: Guilford Press.

Heilbrun, K., Dvoskin, J., Hart, S. D., & McNiel, D. E. (1999). Violence risk communication: Implications for research, policy, and practice. *Health, Risk & Society*, *1*, 91–105.

Heilbrun, K., O'Neill, M. L., Strohman, L. K., Bowman, Q., & Philipson, J. (2000). Expert approaches to communicating violence risk. *Law and Human Behavior*, *24*, 137–148.

Heilbrun, K., Philipson, J., Berman, L., & Warren, J. (1999). Risk communication: Clinicians' reported approaches and perceived values. *Journal of the American Academy of Psychiatry and the Law*, *27*, 397–406.

Helmus, L., Hanson, R. K., & Thornton, D. (2009). Reporting Static-99 in light of new research on recidivism norms. In *The forum* (Vol. 21, pp. 38–45).

Helmus, L., Hanson, R. K., Thornton, D., Babchishin, K. M., & Harris, A. J. R. (2012). Absolute recidivism rates predicted by Static-99R and Static-2002R sex offender risk assessment tools vary across samples: A meta-analysis. *Criminal Justice and Behavior*, *39*, 1148–1171.

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2012). Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights. *Sexual Abuse: A Journal of Research and Treatment*, *24*, 64–101.

Hildebrand, M., De Ruiter, C., & de Vogel, V. (2004). Psychopathy and sexual deviance in treated rapists: Association with sexual and nonsexual recidivism. *Sexual Abuse: A Journal of Research and Treatment*, *16*, 1–24.

Hilton, N. Z., Harris, G. T., & Rice, M. E. (2006). Sixty-six years of research on the clinical versus actuarial prediction of violence. *The Counseling Psychologist*, *34*, 400–409.

Hilton, N. Z., Harris, G. T., Rice, M. E., Houghton, R. E., & Eke, A. W. (2008). An indepth actuarial assessment for wife assault recidivism: The Domestic Violence Risk Appraisal Guide. *Law and Human Behavior*, *32*, 150–163.

Hilton, N. Z., & Simmons, J. L. (2001). The influence of actuarial risk assessment in clinical judgments and tribunal decisions about mentally disordered offenders in maximum security. *Law and Human Behavior*, *25*, 393–408.

Hirschi, T., & Selvin, H. C. (1967). *Delinquency research: An appraisal of analytic methods*. New York, NY: Free Press.

Hodges, E. F. (1971). Crime prevention by the indeterminate sentence law. *American Journal of Psychiatry*, *128*, 291–295.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.

Hoge, R. D., & Andrews, D. A. (2002). *Youth Level of Service/Case Management Inventory users' manual*. North Tonawanda, NY: Multi-Health Systems, Inc.

Holland, T. R., Holt, N., Levi, M., & Beckett, G. E. (1983). Comparison and combination of clinical and statistical predictions of recidivism among adult offenders. *Journal of Applied Psychology*, *68*, 203–211.

Horst, P. (1941). *The prediction of personal adjustment: Survey of logical problems and research techniques, with illustrative application to problems of vocational selection, school success, marriage, and crime*. New York, NY: Social Science Research Council.

Howard, P., Francis, B., Soothill, K., & Humphreys, L. (2009). *OGRS 3: The revised offender group reconviction scale.* Ministry of Justice Research Summary 7/09.

Howell, J. C. (2009). *Preventing and reducing juvenile delinquency: A comprehensive framework* (Second ed.). Sage Publications, Inc.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, *6*, 153–160.

H.R. 4472: 109th Cong., 2nd Sess. (2006). *Adam Walsh Child Protection Act of 2006.* Retrieved from `http://www.gpo.gov`

H.R. 5865: 98th Cong. (1984). *Bail Reform Act of 1984.*

Hubert, L. (1972). A further comment on "N versus N−1". *American Educational Research Journal*, *9*, 323–325.

Hubert, L., & Wainer, H. (2012). *A statistical guide for the ethically perplexed.* Boca Raton, FL: CRC Press.

Huff, R. L. (1936). Is parole prediction a science? *Journal of Criminal Law and Criminology*, *27*, 207–213.

Hunt, R. C., & Wiley, E. D. (1968). Operation Baxstrom after one year. *American Journal of Psychiatry*, *124*, 974–978.

IBM Corportation. (2012). SPSS for Windows CHAID (Version 21.0) [Computer software manual]. Armonk, NY: IBM SPSS Statistics for Windows.

In re Charles Stephenson, 677 Ill.2d. 544. (1977).

In the Matter of the Detention of Richard Hosier, No. 62508-0-I. (Wash. 2010).

In the Matter of the Detention of Scott W. Brooks, 36 P.3d. 1034. (Wash. 2001).

In the Matter of the Personal Restraint of Andre Brigham Young, 857 P.2d. 989. (Wash. 1993).

The Independent: Volume LXXXIV. (1915). New York, NY: Independent Corporation.

Innumerate. (n.d.). *In Merriam-Webster's online dictionary.* Retrieved from `http://www.merriam-webster.com`

International Olympic Committe. (2014, April 19). *IOC regulations on female hyperandrogenism.* Retrieved from `http://www.olympic.org/`

James, D. J., & Glaze, L. E. (2006). *Mental health problems of prison and jail inmates.* Washington, D.C..

James, W., & Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 361–379). University of California Press.

Jan de Bont, Bonnie Curtis, Gerald R. Molen, Walter F. Parkes. (Producers), & Steven Spielberg (Director). (2002). *Minority Report [Motion picture].* United States: Twentieth Century Fox Film Corporation.

Janus, E. S., & Meehl, P. E. (1997). Assessing the legal standard for predictions of dangerousness in sex offender commitment proceedings. *Psychology, Public Policy, and Law*, *3*, 33–64.

Janus, E. S., & Prentky, R. A. (2003). Forensic use of actuarial risk assessment with sex offenders: Accuracy, admissibility and accountability. *American Criminal Law Review*, *40*, 1443-1499.

Jhun, M., & Jeong, H.-C. (2000). Applications of bootstrap methods for categorical data analysis. *Computational Statistics & Data Analysis*, *35*, 83–91.

Jones v. United States, 463 U.S. 354. (1983).

Jurek v. Texas, 428 U.S. 262. (1976).
Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.
Kan. Stats. ch. 59, Art. 29a. (2012).
Kansas v. Crane, 534 U.S. 407. (2002).
Kansas v. Hendricks, 521 U.S. 346. (1997).
Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, *29*, 119–127.
Kelley, T. L. (1923). *Statistical method.* New York, NY: Macmillan Company.
Kirby, B. C. (1954). Parole prediction using multiple correlation. *American Journal of Sociology*, 539–550.
Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin*, *107*, 296–310.
Koerth-Baker, M. (2014, February 17). *Driving under the influence, of marijuana.* Retrieved from `http://www.nytimes.com`
Kolata, G. (2013, November 18). *Flawed gauge for cholesterol risk poses a new challenge for cardiologists.* Retrieved from `http://www.nytimes.com`
Kozol, H. L., Boucher, R. J., & Garofalo, R. (1972). The diagnosis and treatment of dangerousness. *Crime and Delinquency*, *18*, 371–392.
Krauss, D. A., & Sales, B. D. (2001). The effects of clinical and scientific expert testimony on juror decision making in capital sentencing. *Psychology, Public Policy, and Law*, *7*, 267–310.
Kreytak, S. (2010, October). *Longtime expert witness unreliable, court says.* Retrieved from `http://www.statesman.com`
Kroft, S. (Correspondant). (2013, Septemeber 29). Untreated mental illness an imminent danger? [Television series episode]. In Graham Messick, & Coleman Cowan (Producers)*, 60 Minutes.* New York, NY: WCBS.
Kröner, C., Stadtland, C., Eidt, M., & Nedopil, N. (2007). The validity of the Violence Risk Appraisal Guide (VRAG) in predicting criminal recidivism. *Criminal Behaviour and Mental Health*, *17*, 89–100.
Kropp, P. R., & Hart, S. D. (2000). The Spousal Assault Risk Assessment (SARA) Guide: Reliability and validity in adult male offenders. *Law and Human Behavior*, *24*, 101–118.
Kropp, P. R., Hart, S. D., Webster, C. D., & Eaves, D. (1994). *Manual for the Spousal Assault Risk Assessment (SARA) Guide.* Vancouver, Canada: The British Columbia Institute Against Family Violence.
Kropp, P. R., Hart, S. D., Webster, C. D., & Eaves, D. (1999). *The Spousal Assault Risk Assessment: User's guide.* Toronto, Canada: Multi-Health Systems, Inc.
Kumho Tire Co. v. Carmichael, 526 U.S. 137. (1997).
Labi, N. (2011, December 20). *Misfortune teller: A statistics professor says he can predict crime before it occurs.* Retrieved from `http://www.theatlantic.com`
Lamparello, A. (2010). Using cognitive neuroscience to predict future dangerousness. *Columbia Human Rights Law Review*, *42*, 481–539.

Långström, N. (2004). Accuracy of actuarial procedures for assessment of sexual offender recidivism risk may vary across ethnicity. *Sexual Abuse: A Journal of Research and Treatment*, *16*, 107–120.

Langton, C. M., Barbaree, H. E., Seto, M. C., Peacock, E. J., Harkins, L., & Hansen, K. T. (2007). Actuarial assessment of risk for reoffense among adult sex offenders evaluating the predictive accuracy of the Static-2002 and five other instruments. *Criminal Justice and Behavior*, *34*, 37–59.

Lanne, W. F. (1935). Parole prediction as science. *Journal of Criminal Law and Criminology*, *26*, 377–400.

Large, M. M., Ryan, C. J., & Nielssen, O. B. (2010). Helpful and unhelpful risk assessment practices. *Psychiatric Services*, *61*, 530.

Laub, J. H., & Sampson, R. J. (1988). Unraveling families and delinquency: A reanalysis of the Gluecks' data. *Criminology*, *26*, 355–380.

Laub, J. H., & Sampson, R. J. (1991). The Sutherland-Glueck debate: On the sociology of criminological knowledge. *American Journal of Sociology*, *96*, 1402–1440.

Laune, F. F. (1936). *Predicting criminality: Forecasting behavior on parole.* Evanston, IL: Northwestern University Press.

Lave, T. R. (2008). Only yesterday: The rise and fall of twentieth century sexual psychopath laws. *Louisiana Law Review*, *69*, 549–591.

Laves, R. G. (1975). The prediction of dangerousness as a criterion for involuntary civil commitment: Constitutional considerations. *Journal of Psychiatry and Law*, *3*, 291–326.

Leeflang, M. M. G., Bossuyt, P. M. M., & Irwig, L. (2009). Diagnostic test accuracy may vary with prevalence: Implications for evidence-based diagnosis. *Journal of Clinical Epidemiology*, *62*, 5–12.

Levene, K. S., Augimeri, L. K., Pepler, D. J., Walsh, M. M., Webster, C. D., & Koegl, C. J. (2001). *Early assessment risk list for girls: EARL-21G, Version 1 (Consultation edition).* Toronto, Canada: Earlscourt Child and Family Centre.

Lidz, C. W., Mulvey, E. P., & Gardner, W. (1993). The accuracy of predictions of violence to others. *Journal of the American Medical Association*, *269*, 1007–1011.

Lieb, R., & Matson, S. (1998). *Sexual predator commitment laws in the United States: 1998 update.* Olympia, WA: Washington State Institute for Public Policy.

Lilienfeld, S. O., & Jones, M. K. (2008). Allegiance effects in assessment: Unresolved questions, potential explanations, and constructive remedies. *Clinical Psychology: Science and Practice*, *15*, 361–365.

Lipson, M (Producer) & Morris, E (Director). (1988). *The Thin Blue Line* [Motion Picture]. United States of America: Miramax Films.

Litwack, T. R. (1993). On the ethics of dangerousness assessments. *Law and Human Behavior*, *17*, 479–482.

Litwack, T. R. (2001). Actuarial versus clinical assessments of dangerousness. *Psychology, Public Policy, and Law*, *7*, 409–443.

Llanos, M. (2012, December 14). *Authorities ID gunman who killed 27 in elementary school massacre.* Retrieved from `http://usnews.nbcnews.com`

Loeber, R., & Dishion, T. (1983). Early predictors of male delinquency: A review. *Psychological Bulletin*, *94*, 68–99.

Loftus, E. F., & Monahan, J. (1980). Trial by data: Psychological research as legal evidence. *American Psychologist*, *35*, 270–283.

Lombroso, C. (1911). *Crime: Its causes and remedies* (H. P. Horton, Trans.). Boston, MA: Little, Brown, and Company.

Long, L. (2012, December). *Thinking the unthinkable.* Retrieved from `http://anarchistsoccermom.blogspot.com`

Looney, J. W. (2009). Neuroscience's new techniques for evaluating future dangerousness: Are we returning to Lombroso's biological criminality. *UALR Law Review*, *32*, 301–314.

Luborsky, L., Diguer, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., ... Schweizer, E. (1999). The researcher's own therapy allegiances: A "wild card" in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice*, *6*, 95–106.

Luo, M., & McIntire, M. (2013, December 21). *When the right to bear arms includes the mentally ill.* Retrieved from `http://www.nytimes.com`

Lussier, P., Tzoumakis, S., Cale, J., & Amirault, J. (2010). Criminal trajectories of adult sex offenders and the age effect: Examining the dynamic aspect of offending in adulthood. *International Criminal Justice Review*, *20*, 147–168.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences* (Vol. 1). Stanford, CA: Stanford University Press.

Mack, J. L. (1969). MMPI and recidivism. *Journal of Abnormal Psychology*, *74*, 612–614.

MacKenzie, D. L. (2013). First do no harm: A look at correctional policies and programs today. *Journal of Experimental Criminology*, *9*, 1–17.

Maden, A. (2003). Standardised risk assessment: Why all the fuss? *Psychiatric Bulletin*, *27*, 201–204.

Maine Criminal Procedure. (2014, January). *Ch. 15, § 393. Possession of firearms prohibited for certain persons.* Retrieved from `http://www.mainelegislature.org`

Mannheim, H., & Wilkins, L. T. (1955). *Prediction methods in relation to Borstal training* (Vol. 1). London, England: H.M.S.O.

Martin, D. (2011, June). Randall Adams, 61, dies; freed with help of film. *The New York Times*.

Marzban, C. (2004). The ROC curve and the area under it as performance measures. *Weather and Forecasting*, *19*, 1106–1114.

MATLAB. (2013). Version 8.2 (R2013b) [Computer software manual]. Natick, MA: The MathWorks Inc.

Matthew v. Nelson, 461 F. Supp. 707. (1978).

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York, NY: Houghton Mifflin Harcourt.

McCusker, P. J. (2007). Issues regarding the clinical use of the Classification of Violence Risk (COVR) assessment instrument. *International Journal of Offender Therapy and Comparative Criminology*, *51*, 676–685.

McDermott, B. E., Dualan, I. V., & Scott, C. L. (2011). The predictive ability of the classification of violence risk (COVR) in a forensic psychiatric hospital. *Psychiatric Services*, *62*, 430–433.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology*, *50*, 215–241.

McMillan, D., Hastings, R. P., & Coldwell, J. (2004). Clinical and actuarial prediction of physical violence in a forensic intellectual disability hospital: A longitudinal study. *Journal of Applied Research in Intellectual Disabilities*, *17*, 255–265.

McNiel, D. E., & Binder, R. L. (1994a). The relationship between acute psychiatric symptoms, diagnosis, and short-term risk of violence. *Hospital and Community Psychiatry*, *45*, 133–137.

McNiel, D. E., & Binder, R. L. (1994b). Screening for risk of inpatient violence. *Law and Human Behavior*, *18*, 579–586.

McNiel, D. E., Gregory, A. L., Lam, J. N., Binder, R. L., & Sullivan, G. R. (2003). Utility of decision support tools for assessing acute risk of violence. *Journal of Consulting and Clinical Psychology*, *71*, 945–953.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology*, *4*, 268–273.

Meehl, P. E., & Rosen, A. (1955). Antecedent probabilitiy and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194–215.

Megargee, E. I. (1976). The prediction of dangerous behavior. *Criminal Justice and Behavior*, *3*, 3–22.

Melamed, Y., Bauer, A., Kalian, M., Rosca, P., & Mester, R. (2011). Assessing the risk of violent behavior before issuing a license to carry a handgun. *Journal of the American Academy of Psychiatry and the Law*, *39*, 543–548.

Memmott, M. (2014, April 10). *School stabbing suspect was 'nice young boy,' attorney says.* Retrieved from `http://www.npr.org`

Mental Health Act, C. 20. (1983).

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*, 283–298.

Millard v. Harris, 406 F. 2d 964. (1968).

Miller, A. B., Wall, C., Baines, C. J., Sun, P., To, T., & Narod, S. A. (2014). Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: Randomised screening trial. *British Medical Journal*, *348*, 1–10.

Miller, C. S., Kimonis, E. R., Otto, R. K., Kline, S. M., & Wasserman, A. L. (2012). Reliability of risk assessment measures used in sexually violent predator proceedings. *Psychological Assessment*, *24*, 944–953.

Mills, J. F., Jones, M. N., & Kroner, D. G. (2005). An examination of the generalizability of the LSI-R and VRAG probability bins. *Criminal Justice and Behavior*, *32*, 565–585.

Mitchell, J. N. (1969). Bail reform and the constitutionality of pretrial detention. *Virginia Law Review*, *56*, 1223–1242.

Moeller, K. E., Lee, K. C., & Kissack, J. C. (2008). Urine drug screening: Practical guide for clinicians. In *Mayo clinic proceedings* (Vol. 83, pp. 66–76).

Monachesi, E. D. (1932). *Prediction factors in probation: A study of 1515 probation cases of Ramsey County, Minnesota for the years 1923-1925* (Vol. 2). Hanover, NH: The Sociological Press.

Monachesi, E. D. (1939). Can we predict probation outcomes? *Federal Probation*, *3*, 15–18.

Monachesi, E. D. (1945). A comparison of predicted with actual results of probation. *American Sociological Review*, *10*, 26–31.

Monachesi, E. D. (1948). Some personality characteristics of delinquents and non-delinquents. *Journal of Criminal Law and Criminology*, *38*, 487–500.

Monachesi, E. D. (1950a). Personality characteristics and socio-economic status of delinquents and non-delinquents. *Journal of Criminal Law and Criminology*, *40*, 570–583.

Monachesi, E. D. (1950b). Personality characteristics of institutionalized and non-institutionalized male delinquents. *Journal of Criminal Law and Criminology*, *41*, 487–500.

Monahan, J. (1973). Dangerous offenders a critique of Kozol et al. *Crime and Delinquency*, *19*, 418–420.

Monahan, J. (1977). Strategies for an empirical analysis of the prediction of violence in emergency civil commitment. *Law and Human Behavior*, *1*, 363–371.

Monahan, J. (1981). *Predicting violent behavior: An assessment of clinical techniques*. Beverly Hills, CA: Sage Publications, Inc.

Monahan, J. (1984). The prediction of violent behavior: Toward a second generation of theory and policy. *American Journal of Psychiatry*, *141*, 10–15.

Monahan, J. (1988). Risk assessment of violence among the mentally disordered: Generating useful knowledge. *International Journal of Law and Psychiatry*.

Monahan, J. (1992). Mental disorder and violent behavior: Perceptions and evidence. *American Psychologist*, *47*, 511–521.

Monahan, J. (1993). Limiting therapist exposure to *Tarasoff* liability: Guidelines for risk containment. *American Psychologist*, *48*, 242–250.

Monahan, J. (1996). Violence prediction the past twenty and the next twenty years. *Criminal Justice and Behavior*, *23*, 107–120.

Monahan, J. (2006). A jurisprudence of risk assessment: Forecasting harm among prisoners, predators, and patients. *Virginia Law Review*, *92*, 391–435.

Monahan, J. (2007). Statistical literacy: A prerequisite for evidence-based medicine. *Psychological Science in the Public Interest*, *8*, i–ii.

Monahan, J. (2012). The individual risk assessment of terrorism. *Psychology, Public Policy, and Law*, *18*, 167–205.

Monahan, J. (2013). Bioprediction, biomarkers, and bad behavior: Scientific, legal, and ethical challenges. In I. Singh, W. P. Sinnott-Armstrong, & J. Savulescu (Eds.), (pp. 57–76). New York, NY: Oxford University Press.

Monahan, J., & Cummings, L. (1974). Prediction of dangerousness as a function of its perceived consequences. *Journal of Criminal Justice*, *2*, 239–242.

Monahan, J., & Cummings, L. (1975). Social policy implications of the inability to predict violence. *Journal of Social Issues*, *31*, 153–164.

Monahan, J., Heilbrun, K., Silver, E., Nabors, E., Bone, J., & Slovic, P. (2002). Communicating violence risk: Frequency formats, vivid outcomes, and forensic settings. *International Journal of Forensic Mental Health*, *1*, 121–126.

Monahan, J., & Silver, E. (2003). Judicial decision thresholds for violence risk management. *International Journal of Forensic Mental Health*, *2*, 1–6.

Monahan, J., & Steadman, H. J. (1996). Violent storms and violent people: How meteorology can inform risk communication in mental health law. *American Psychologist*, *51*, 931–938.

Monahan, J., Steadman, H. J., Appelbaum, P. S., Grisso, T., Mulvey, E. P., Roth, L. H., ... Silver, E. (2006). The Classification of Violence Risk. *Behavioral Sciences and the Law*, *24*, 721–730.

Monahan, J., Steadman, H. J., Robbins, P. C., Appelbaum, P. S., Banks, S., Grisso, T., ... Silver, E. (2005). An actuarial model of violence risk assessment for persons with mental disorders. *Psychiatric Services*, *56*, 810–815.

Monahan, J., Steadman, H. J., Robbins, P. C., Silver, E., Appelbaum, P. S., Grisso, T., ... Roth, L. H. (2000). Developing a clinically useful actuarial tool for assessing violence risk. *The British Journal of Psychiatry*, *176*, 312–319.

Monahan, J., Steadman, H. J., Silver, E., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., ... Banks, S. (2001). *Rethinking risk assessment: The MacArthur study of mental disorder and violence*. New York, NY: Oxford University Press.

Monahan, J., & Wexler, D. B. (1978). A definite maybe: Proof and probability in civil commitment. *Law and Human Behavior*, *2*, 37–42.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, *58*, 415–434.

Morris, G. H. (1967). The confusion of confinement syndrome: An analysis of the confinement of mentally ill criminals and ex-criminals by the Department of Correction of the State of New York. *Buffalo Law Review*, *17*, 651–699.

Mossman, D. (1994a). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology*, *62*, 783–792.

Mossman, D. (1994b). Further comments on portraying the accuracy of violence predictions. *Law and Human Behavior*, *18*, 587–593.

Mossman, D. (2006a). Another look at interpreting risk categories. *Sexual Abuse: A Journal of Research and Treatment*, *18*, 41–63.

Mossman, D. (2006b). Critique of pure risk assessment or, Kant meets *Tarasoff*. *University of Cincinnati Law Review*, *75*, 523–609.

Mossman, D. (2008). Analyzing the performance of risk assessment instruments: A response to Vrieze and Grove (2007). *Law and Human Behavior*, *32*, 279–291.

Mossman, D. (2013). Evaluating risk assessments using receiver operating characteristic analysis: Rationale, advantages, insights, and limitations. *Behavioral Sciences and the Law*, *31*, 23–39.

Mossman, D., Schwartz, A. H., & Elam, E. R. (2011). Risky business versus overt acts: What relevance do actuarial, probabilistic risk assessments have for judicial decisions on involuntary psychiatric hospitalization. *Houston Journal of Health Law and Policy*, 365–453.

Mossman, D., & Sellke, T. (2007). Avoiding errors about 'margins of error'. *The British Journal of Psychiatry*, *191*, 561.

Murphy, A. H. (1972a). Scalar and vector partitions of the probability score: part I. Two-state situation. *Journal of Applied Meteorology*, *11*, 273–282.

Murphy, A. H. (1972b). Scalar and vector partitions of the probability score: part II. N-state situation. *Journal of Applied Meteorology*, *11*, 1183–1192.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, *12*, 595–600.

Murrie, D. C., Boccaccini, M. T., Johnson, J. T., & Janke, C. (2008). Does interrater (dis)agreement on Psychopathy Checklist scores in sexually violent predator trials suggest partisan allegiance in forensic evaluations? *Law and Human Behavior*, *32*, 352–362.

Nadelhoffer, T., Bibas, S., Grafton, S., Kiehl, K. A., Mansfield, A., Sinnott-Armstrong, W., & Gazzaniga, M. (2012). Neuroprediction, violence, and the law: Setting the stage. *Neuroethics*, *5*, 67–99.

Nadelhoffer, T., & Sinnott-Armstrong, W. (2012). Neurolaw and neuroprediction: Potential promises and perils. *Philosophy Compass*, *7*, 631–642.

National Public Radio. (2013a, May 16). *Analyzing the language of suicide notes to help save lives.* Retrieved from `http://www.npr.org`

National Public Radio. (2013b, June 7). *Whole genome scans could reveal too much.* Retrieved from `http://www.npr.org`

Negro v. Dickens, 22 A.D. 2d, 406. (1965).

Neller, D. J., & Frederick, R. I. (2013). Classification accuracy of actuarial risk assessment instruments. *Behavioral Sciences and the Law*, *153*, 141–153.

Neuilly, M.-A., Zgoba, K. M., Tita, G. E., & Lee, S. S. (2011). Predicting recidivism in homicide offenders using classification tree analysis. *Homicide Studies*, *15*, 154–176.

New Jersey Code of Criminal Justice. (2014, January). *Title 2C:7-1 to 11.* Retrieved from `http://lis.njleg.state.nj.us`

New York Civil Liberties Union (NYCLU). (2014, January). *Stop-and-frisk data.* Retrieved from `http://www.nyclu.org`

New York Criminal Procedure. (2014, January). *Art. 140 § 140.50 Temporary Questioning of Persons in Public Places; Search for Weapons.* Retrieved from `http://law.onecle.com`

New York Family Court Law. (n.d.). *Art. 3, juvenile delinqency.*

New York State. (2013, May). *NY SAFE Act.* Retrieved from `http://www.governor.ny.gov`

Newman, K. S., Fox, C., Roth, W., Mehta, J., & Harding, D. (2005). *Rampage: The social roots of school shootings.* New York, NY: Basic Books.

N.H. Stats. (2014a, January). *Title LII: Proceedings in court.* Retrieved from `http://www.gencourt.state.nh.us`

N.H. Stats. (2014b, January). *Title X: Public health.* Retrieved from `http://www.gencourt.state.nh.us`

Nicholls, T. L., Ogloff, J. R. P., & Douglas, K. S. (2004). Assessing risk for violence among male and female civil psychiatric patients: The HCR-20, PCL:SV, and VSC. *Behavioral Sciences and the Law*, *22*, 127–158.

Nixon, R. (1969, January 31). *Statement outlining actions and recommendations for the District of Columbia.* The American Presidency Project. Retrieved from `http://www.presidency.ucsb.edu`

Nixon, R. (1970, July 29, 1970). *Remarks on signing the District of Columbia Court Reform and Criminal Procedure Act of 1970.* The American Presidency Project. Retrieved from `http://www.presidency.ucsb.edu`

Nuffield, J. (1982). *Parole decision-making in Canada: Research towards decision guidelines.* Ottawa, Canada: Solicitor General of Canada.

N.Y. Mental Hygiene Law, § 9.60. (1999).

Ohlin, L. E., & Duncan, O. D. (1949). The efficiency of prediction in criminology. *American Journal of Sociology*, 441–452.

Ohlin, L. E., & Lawrence, R. A. (1952). A comparison of alternative methods of parole prediction. *American Sociological Review*, *17*, 268–274.

Olver, M. E. (2003). *The development and validation of the Violence Risk Scale: Sexual Offender version (VRS:SO) and its relationship to psychopathy and treatment attrition* (Unpublished doctoral dissertation). University of Saskatchewan.

Olver, M. E., Wong, S. C. P., Nicholaichuk, T., & Gordon, A. (2007). The validity and reliability of the Violence Risk Scale-Sexual Offender version: Assessing sex offender risk and evaluating therapeutic change. *Psychological Assessment*, *19*, 318–329.

Otto, R. K. (1992). Prediction of dangerous behavior: A review and analysis of "second-generation" research. *Forensic Reports*, *5*, 103–134.

Otto, R. K., & Petrila, J. (2002). Admissibility of testimony based on actuarial scales in sex offender commitments: A reply to Doren. *Sex Offender Law Report*, *3*, 1, 14–16.

Overall, J. E., & Gorham, D. R. (1962). The Brief Psychiatric Rating Scale. *Psychological Reports*, *10*, 799–812.

Overholser v. Russell, 283 F. 2d. 195. (1960).

Pearson, K. (1900a). On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *195*, 1–47.

Pearson, K. (1900b). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, *50*, 157–175.

People v. Murtishaw 29 Cal. 3d 733. (1981).

People v. Poddar, 518 P. 2d 342. (1974).

People v. Superior Court (Ghilotti), 119 Cal. Rptr. 2d 1. (2002).

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction.* Oxford, Englad: Oxford University Press.

Pescosolido, B. A., Monahan, J., Link, B. G., Stueve, A., & Kikuzawa, S. (1999). The public's view of the competence, dangerousness, and need for legal coercion of persons with mental health problems. *American Journal of Public Health*, *89*, 1339–1345.

Pierson, J. (2011). Construing *Crane*: Examining how state courts have applied its lack-of-control standard. *University of Pennsylvania Law Review*, *160*, 1527–1559.

Pintner, R., & Reamer, J. (1918). Mental ability and future success of delinquent girls. *Journal of Delinquency*, *3*, 74–79.

Plumer, B. (2012a, December 17). *Graph of the day: Perhaps mass shootings aren't becoming more common.* Retrieved from http://www.washingtonpost.com

Plumer, B. (2012b, December 14). *Why are mass shootings increasing even while gun violence is decreasing?* Retrieved from http://www.washingtonpost.com

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, *1*, 125–126.

Povoledo, E., & Fountain, H. (2012, October 22). *Italy order jail terms for 7 who didn't wanr of deadly earthquake.* Retrieved from http://www.nytimes.com

Powers, E., & Witmer, H. (1951). *An experiment in the prevention of delinquency: The Cambridge-Sommerville Youth Study.* New York, NY: Columbia University Press.

PredPol. (2014, January). *PredPol: Predicitive policing in a box.* Retrieved from http://www.predpol.com

Prescott, M. (2009). Invasion of the body snatchers: Civil commitment after Adam Walsh. *University of Pittsburgh Law Review*, *71*, 839–884.

Prochaska v. Brinegar, 102 N.W. 2d 870. (1979).

Purifoy v. Watters, 197 Wis 2d. 279. (2010).

Quinsey, V. L. (1979). Assessments of the dangerousness of mental patients held in maximum security. *International Journal of Law and Psychiatry*, *2*, 389–406.

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (1999). *Violent offenders: Appraising and managing risk* (First ed.). Washington, D.C.: American Psychological Association.

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (2006). *Violent offenders: Appraising and managing risk* (Second ed.). Washington, D.C.: American Psychological Association.

Quinsey, V. L., Rice, M. E., & Harris, G. T. (1995). Actuarial prediction of sexual recidivism. *Journal of Interpersonal Violence*, *10*, 85–105.

R Core Team. (2012). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Raine, A. (2013). *The anatomy of violence: The biological roots of crime.* New York, NY: Random House, Inc.

Ralston, C. A., & Epperson, D. L. (2013). Predictive validity of adult risk assessment tools with juveniles who offended sexually. *Psychological Assessment*, *25*, 905–916.

Ransohoff, D. F., & Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine*, *299*, 926–930.

Rauh, C. S., & Silbert, E. J. (1970). Criminal law and procedure: D.C. Court Reform and Criminal Procedure Act of 1970. *The American University Law Review*, *20*, 252–341.

Rayman, G. (2010, May 4). *The NYPD tapes: Inside Bed-Stuy's 81st Precinct.* Retrieved from http://www.villagevoice.com

Reid, W. H. (2001). Psychiatry and the death penalty. *Journal of Psychiatric Practice*, *7*, 216–219.

Reidy, T. J., Cunningham, M. D., & Sorensen, J. R. (2001). From death to life prison behavior of former death row inmates in Indiana. *Criminal Justice and Behavior*, *28*, 62–82.

Reiss, A. J., Jr. (1951). Unraveling juvenile delinquency. II. An appraisal of the research methods. *American Journal of Sociology*, *57*, 115–120.

Rettenberger, M., Boer, D. P., & Eher, R. (2011). The predictive accuracy of risk factors in the Sexual Violence Risk–20 (SVR-20). *Criminal Justice and Behavior*, *38*, 1009–1027.

Revised Code of Washington, 71.09.020(7). (2013).

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology*, *63*, 737–748.

Rice, M. E., & Harris, G. T. (1997). Cross-validation and extension of the Violence Risk Appraisal Guide for child molesters and rapists. *Law and Human Behavior*, *21*, 231–241.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's *d*, and *r*. *Law and Human Behavior*, *29*, 615–620.

Rice, M. E., Harris, G. T., & Hilton, N. Z. (2010). Handbook of violence risk assessment. In R. K. Otto & K. S. Douglas (Eds.), (pp. 99–119). New York, NY: Routledge.

Rice, M. E., Harris, G. T., & Lang, C. (2013). Validation of and revision to the VRAG and SORAG: The Violence Risk Appraisal Guide–Revised (VRAG-R). *Psychological Assessment*, *25*, 951–965.

Rice, S. A. (1928). Some inherent difficulties in the method of prediction by classification. *Social Forces*, *7*, 554–558.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*, 1–25.

Roberts, C. F., Doren, D. M., & Thornton, D. (2002). Dimensions associated with assessments of sex offender recidivism risk. *Criminal Justice and Behavior*, *29*, 569–589.

Roberts, R. S., Spitzer, W. O., Delmore, T., & Sackett, D. L. (1978). An empirical demonstration of Berkson's bias. *Journal of Chronic Diseases*, *31*, 119–128.

Ronayne, J. A. (1964). The right to investigate and New York's Stop and Frisk Law. *Fordham Law Review*, 211–238.

Rose, G. (1966). Trends in the use of prediction. *The Howard Journal of Criminal Justice*, *12*, 26–33.

Rosenbaum, R. (1990, May). Travels with Dr. Death. *Vanity Fair*, 141–166.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rubin, B. (1972). Prediction of dangerousness in mentally ill criminals. *Archives of General Psychiatry*, *27*, 397–407.

Rubin, J. (2010, August 21). *Stopping crime before it starts.* Retrieved from `http://articles.latimes.com`

Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*, *32*, 51–63.

Saldano v. Roach, 363 F. 3d 545. (2004).
Saldano v. State, S.W. 3d 873. (2002).
Sallinger, R. (2012, August 21). *James Holmes saw three mental health professionals before shooting.* Retrieved from http://www.cbsnews.com
Sanders, F. (1958). The evaluation of subjective probability forecasts. Techinal Report No. 5, Contract AF 19(604)-1305, Department of Meteorology, MIT.
Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*, *2*, 191–201.
Sarbin, T. R. (1943). A contribution to the study of actuarial and individual methods of prediction. *American Journal of Sociology*, 593–602.
Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin*, *66*, 178–200.
Schall v. Martin, 467 U.S. 253. (1984).
Scherr, A. (2003). *Daubert* & danger: The fit of expert predictions in civil commitments. *Hastings Law Journal*, *55*, 1–90.
Schlager, M. D., & Simourd, D. J. (2007). Validity of the Level of Service InventoryRevised (LSI-R) among African American and Hispanic male offenders. *Criminal Justice and Behavior*, *34*, 545–554.
Schlesinger, S. E. (1978). The prediction of dangerousness in juveniles: A replication. *Crime and Delinquency*, *24*, 40–48.
Schneps, L., & Colmez, C. (2013). *Math on trial: How numbers get used and abused in the courtroom.* Philadelphia, PA: Basic Books.
Schuessler, K. F., & Cressey, D. R. (1950). Personality characteristics of criminals. *American Journal of Sociology*, 476–484.
Schwarz, A. (2014, April 11). *Idea of new attention disorder spurs research, and debate.* Retrieved from http://www.nytimes.com
Schwarz, N., Strack, F., Hilton, D., & Naderer, G. (1991). Base rates, representativeness, and the logic of conversation: The contextual relevance of "irrelevant" information. *Social Cognition*, *9*, 67–84.
Scott, P. D. (1977). Assessing dangerousness in criminals. *The British Journal of Psychiatry*, *131*, 127–142.
Scurich, N., & John, R. S. (2011). The effect of framing actuarial risk probabilities on involuntary civil commitment decisions. *Law and Human Behavior*, *35*, 83–91.
Scurich, N., & John, R. S. (2012). A Bayesian approach to the group versus individual prediction controversy in actuarial risk assessment. *Law and Human Behavior*, *36*, 237–246.
Scurich, N., Monahan, J., & John, R. S. (2012). Innumeracy and unpacking: Bridging the nomothetic/idiographic divide in violence risk assessment. *Law and Human Behavior*, *36*, 548–554.
Sengupta, S. (2013, June 19). *In hot pursuit of numbers to ward off crime.* Retrieved from http://bits.blogs.nytimes.com
Seto, M. C. (2005). Is more better? Combining actuarial risk scales to predict recidivism among adult sex offenders. *Psychological Assessment*, *17*, 156–167.

Shah, S. A. (1978). Dangerousness: A paradigm for exploring some issues in law and psychology. *American Psychologist*, *33*, 224–238.

Shapiro, J. (2014, April 3). *Shooting unfairly links violence with mental illness—again.* Retrieved from `http://www.npr.org`

Shaw, S. H. (1973). The dangerousness of dangerousness. *Medicine, Science, and the Law*, *13*, 269–271.

Silver, E. (2006). Understanding the relationship between mental disorder and violence: The need for a criminological perspective. *Law and Human Behavior*, *30*, 685–706.

Silver, E., Smith, W. R., & Banks, S. (2000). Constructing actuarial devices for predicting recidivism: A comparison of methods. *Criminal Justice and Behavior*, *27*, 733–764.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail—but some don't.* New York, NY: The Penguin Press.

Simon, R. J., & Mahan, L. (1971). Quantifying burdens of proof: A view from the bench, the jury, and the classroom. *Law and Society Review*, *5*, 319–330.

Singh, J. P. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences and the Law*, *31*, 8–22.

Singh, J. P., Desmarais, S. L., & Van Dorn, R. A. (2013). Measurement of predictive validity in violence risk assessment studies: A second-order systematic review. *Behavioral Sciences an the Law*, *31*, 55–73.

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review*, *31*, 499–513.

Singh, J. P., Grann, M., & Fazel, S. (2013). Authorship bias in violence risk assessment? A systematic review and meta-analysis. *PloS One*, *8*, 1–8.

Singh, J. P., Serper, M., Reinharth, J., & Fazel, S. (2011). Structured assessment of violence risk in schizophrenia and other psychiatric disorders: A systematic review of the validity, reliability, and item content of 10 available instruments. *Schizophrenia Bulletin*, *37*, 899–912.

Sjöstedt, G., & Grann, M. (2002). Risk assessment: What is being predicted by actuarial prediction instruments? *International Journal of Forensic Mental Health*, *1*, 179–183.

Sjöstedt, G., & Långström, N. (2001). Actuarial assessment of sex offender recidivism risk: A cross-validation of the RRASOR and the Static-99 in Sweden. *Law and Human Behavior*, *25*, 629–645.

Sjöstedt, G., & Långström, N. (2002). Assessment of risk for criminal recidivism among rapists: A comparison of four different measures. *Psychology, Crime and Law*, *8*, 25–40.

Skeem, J. L., Manchak, S. M., Lidz, C. W., & Mulvey, E. P. (2013). The utility of patients' self-perceptions of violence risk: Consider asking the person who may know best. *Psychiatric Services*.

Skeem, J. L., & Monahan, J. (2011). Current directions in violence risk assessment. *Current Directions in Psychological Science*, *20*, 38–42.

Skeem, J. L., & Mulvey, E. P. (2002). Care of the mentally disordered offender in the community. In A. Buchanan (Ed.), (pp. 111–142). New York, NY: Oxford Press.

Skeem, J. L., Schubert, C., Odgers, C., Mulvey, E. P., Gardner, W., & Lidz, C. W. (2006). Psychiatric symptoms and community violence among high-risk patients: A test of the relationship at the weekly level. *Journal of Consulting and Clinical Psychology*, *74*, 967–979.

Slobogin, C. (1984). Dangerousness and expertise. *University of Pennsylvania Law Review*, *133*, 97–174.

Slobogin, C. (2006). Dangerousness and expertise redux. *Emory Law Journal*, *56*, 275–326.

Slobogin, C. (2009). Criminal law conversations. In (pp. 67–74). New York, NY: Oxford University Press.

Slovic, P., & Monahan, J. (1995). Probability, danger, and coercion: A study of risk perception and decision making in mental health law. *Law and Human Behavior*, *19*, 49–65.

Slovic, P., Monahan, J., & MacGregor, D. G. (2000). Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior*, *24*, 271–296.

Snowden, R. J., Gray, N. S., Taylor, J., & Fitzgerald, S. (2009). Assessing risk of future violence among forensic psychiatric inpatients with the classification of violence risk (COVR). *Psychiatric Services*, *60*, 1522–1526.

Snowden, R. J., Gray, N. S., Taylor, J., & MacCulloch, M. J. (2007). Actuarial prediction of violent recidivism in mentally disordered offenders. *Psychological Medicine*, *37*, 1539–1550.

Song, L., & Lieb, R. (1995). *Washington State sex offenders: Overview of recidivism studies.* Olympia, WA: Washington State Institute for Public Policy.

Sorensen, J. R., & Pilgrim, R. L. (1999). An actuarial risk assessment of violence posed by capital murder defendants. *Journal Criminal Law and Criminology*, *90*, 1251–1270.

Soumitra. (2014, January). *Plot summary for Minority Report (2002)* [Review of *Minority Report*, produced by Jan de Bont, Bonnie Curtis, Gerald R. Molen, and Walter F. Parkes, directed by Steven Spielberg, 2002]. Retrieved from `http://www.imdb.com`

SPSS, I. (1993). SPSS for Windows (Release 6.0) [Computer software manual]. Chicago, IL: Author

Sreenivasan, S., Weinberger, L. E., Frances, A., & Cusworth-Walker, S. (2010). Alice in actuarial-land: Through the looking glass of changing Static-99 norms. *Journal of the American Academy of Psychiatry and the Law*, *38*, 400–406.

Sreenivasan, S., Weinberger, L. E., & Garrick, T. (2003). Expert testimony in sexually violent predator commitments: Conceptualizing legal standards of "mental disorder" and "likely to reoffend". *Journal of the American Academy of Psychiatry and the Law Online*, *31*, 471–485.

State of Connecticut Division of Criminal Justice. (2013). *Report of the state's attorney for the judicial district of Danbury on the shootings at Sandy Hook Elementary School and 36 Yogananda street, Newtown, Connecticut on December 14, 2012.*

State of New Hampshire v. William Ploof, No. 07-E-0238. (2009).

State v. Curiel, 227 Wis. 2d. 389. (1999).

State v. Krol, 344 A. 2d. 289. (1975).

State v. Post, 197 Wis. 2d 279. (1995).
State v. Randall, 192 Wis. 2d 800. (1995).
Static-99. (2013, February). *Static-99 documents.* Retrieved from www.static99.org
Steadman, H. J. (1973). Some evidence on the inadequacy of the concept and determination of dangerousness in law and psychiatry. *Journal of Psychiatry and Law*, *1*, 409–426.
Steadman, H. J., & Cocozza, J. J. (1974). *Careers of the criminally insane: Excessive social control of deviance.* Lexington, MA: Lexington Books.
Steadman, H. J., Silver, E., Monahan, J., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., ... Banks, S. (2000). A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior*, *24*, 83–100.
Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 197–206). University of California Press.
Stein, R. (2014, February 26). *Blood test provides more accurate prenatal testing for down syndrome.* Retrieved from http://www.npr.org
Stone, A. A. (1975). *Mental health and law: A system in transition.* Rockville, MD: National Institute of Mental Health, Center for Studies of Crime and Delinquency.
Strasser, A.-R. (2014, January 2). *Mass shootings are becoming more frequent.* Retrieved from http://thinkprogress.org
Stuart, H. (2003). Violence and mental illness: An overview. *World Psychiatry*, *2*, 121–124.
Sturup, J., Kristiansson, M., & Lindqvist, P. (2011). Violent behaviour by general psychiatric patients in Sweden: Validation of Classification of Violence Risk (COVR) software. *Psychiatry Research*, *188*, 161–165.
Sutherland, E. H. (1937). Review of Gluecks' *Later Criminal Careers. Harvard Law Review*, *51*, 184–186.
Swanson, J. W., Holzer III, C. E., Ganju, V. K., & Jono, R. T. (1990). Violence and psychiatric disorder in the community: Evidence from the Epidemiologic Catchment Area surveys. *Hospital and Community Psychiatry*, *41*, 761–770.
Sweetland, J. P. (1973). *"Illusory correlation" and the estimation of "dangerous" behavior.* (Unpublished doctoral dissertation). Indiana University.
Swets, J. A., Dawes, R. M., & Monahan, J. (2000a). Better decisions through science. *Scientific American*, *283*, 82–87.
Swets, J. A., Dawes, R. M., & Monahan, J. (2000b). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26.
Szmukler, G. (2001). Violence risk prediction in practice. *The British Journal of Psychiatry*, *178*, 84–85.
Szmukler, G. (2003). Risk assessment: 'Numbers' and 'values'. *Psychiatric Bulletin*, *27*, 205–207.
Szmukler, G., Everitt, B., & Leese, M. (2012). Risk assessment and receiver operating characteristic curves. *Psychological Medicine*, *42*, 895–898.
Tangney, J. P., Stuewig, J., & Martinez, A. G. (2014). Two faces of shame the roles of shame and guilt in predicting recidivism. *Psychological Science*, 1–7.
Tarasoff v. Regents of the University of California, 17 Cal. 3d 425. (1976).

Tarling, R. (1982). Comparison of measures of predictive power. *Educational and Psychological Measurement*, *42*, 479–487.

Taylor, K. (2012, June 10). *Stop-and-frisk policy 'saves lives,' mayor tells black congregation.* Retrieved from http://www.nytimes.com

Taylor, R. (1999). *Predicting reconvictions for sexual and violent offences using the revised Offender Group Reconviction Scale,* Research Findings*, no. 104.* London, England: Home Office.

Terry v. Ohio, 392 U.S. 1. (1968).

Texas Code of Criminal Procedure. (2013, May). *Art. 37.071. Procedure in Capital Case.* Retrieved from http://www.statutes.legis.state.tx.us

Texas Code of Criminal Procedure. (2014, January). *Art. 37.07. Verdict Must Be General; Separate Hearing on Proper Punishment.* Retrieved from http://www.statutes.legis.state.tx.us

Texas Defender Service. (2004). *Deadly speculation: Misleading Texas capital juries with false predictions of future dangerousness.* Houston, TX: Author.

The Associated Press. (2013, October 21). *Eric Holder: Number of mass shootings tripled.* Retrieved from www.politico.com

The Association for the Treatment of Sexual Abusers (ATSA). (2014, January). *Civil commitment of sexually violent predators.* Retrieved from http://www.atsa.com

The New York Times. (2014, April 6). *The 'superpredator' scare [Video file].* Retrieved from http://www.nytimes.com

Thompson, R. E. (1952). A validation of the Glueck Social Prediction Scale for proneness to delinquency. *The Journal of Criminal Law, Criminology, and Police Science*, *43*, 451–470.

Thompson, R. E. (1957). Further validation of the Glueck social prediction table for identifying potential delinquents. *Journal of Criminal Law, Criminology, and Police Science*, *48*, 175–184.

Thompson, W. C., & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior*, *11*, 167–187.

Thornton, D. (2007). *Scoring guide for risk matrix 2000.9/SVC.* Retrieved from http://www.bhamlive1.bham.ac.uk

Tibbitts, C. (1931). Success or failure on parole can be predicted: A study of the records of 3,000 youths paroled from the Illinois State Reformatory. *Journal of Criminal Law and Criminology (1931-1951)*, *22*, 11–50.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267–288.

Tierney, J. (2014, March 23). *At airports, a misplaced faith in body language.* Retrieved from http://www.nytimes.com

Tikhonov, A. N. (1943). On the stability of inverse problems. *Doklady Akademii Nauk SSSR [Proceeding of the USSR Academy of Sciences]*, *39*, 195–198.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics*, *4*, 1035–1038.

Torrey, E. F., Kennard, A. D., Eslinger, D., Lamb, R., & Pavle, J. (2010). *More mentally ill persons are in jails and prisons than hospitals: A survey of the states.* Arlington, VA: Treatment Advocacy Center.

Treatment Advocacy Center. (2013, December). *Preventable tragedies.* Treatment Advocacy Center. Retrieved from `http://www.treatmentadvocacycenter.org`

Tribe, L. H. (1970). An ounce of detention: Preventive justice in the world of John Mitchell. *Virginia Law Review*, 371–407.

Tribe, L. H. (1971). Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review*, *84*, 1329–1393.

Tukey, J. W. (1958). Bias and confidence in not quite large samples [abstract]. *Annals of Mathematical Statistics*, *29*, 614.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

Underwood, B. D. (1977). The thumb on the scales of justice: Burdens of persuasion in criminal cases. *The Yale Law Journal*, *86*, 1299–1348.

Underwood, B. D. (1979). Law and the crystal ball: Predicting behavior with statistical inference and individualized judgment. *The Yale Law Journal*, *88*, 1408–1448.

United States v. Abregana, 574 F. Supp. 2d 1145. (2008).

United States v. Comstock, 130 U.S. 126. (2010).

United States v. Edwards, 777 F. Supp. 2d 985. (2011).

United States v. Fatico, 458 F. Supp. 388. (1978).

United States v. Rehlander, 666 F. 3d 45. (2012).

United States v. Salerno, 481 U.S. 739. (1987).

United States v. Wooden, 693 F. 3d 440. (2012).

U.S. Const. amend. IV. (n.d.).

U.S. Const. amend. V. (n.d.).

U.S. Const. amend. VI. (n.d.).

U.S. Const. amend. VIII. (n.d.).

U.S. Const. amend. XIV, § 1. (n.d.).

U.S. Department of Homeland Security. (2008). *Privacy Impact Assessment for the Future Attribute Screening Technology (FAST) Project.*

Vecchio, T. J. (1966). Predictive value of a single diagnostic test in unselected populations. *New England Journal of Medicine*, *274*, 1171–1173.

Vitacco, M. J., Erickson, S. K., Kurus, S., & Apple, B. N. (2012). The role of the Violence Risk Appraisal Guide and Historical, Clinical, Risk-20 in US courts: A case law survey. *Psychology, Public Policy, and Law*, *18*, 361–391.

Vlahos, J. (2011). The department of pre-crime. *Scientific American*, *306*, 62–67.

Vold, G. B. (1931). *Prediction methods and parole: A study of factors involved in the violation or non-violation of parole in a group of Minnesota adult males.* Hanover, NH: The Sociological Press.

Vrieze, S. I., & Grove, W. M. (2008). Predicting sex offender recidivism. I. Correcting for item overselection and accuracy overestimation in scale development. II. Sampling error-induced attenuation of predictive validity over base rate information. *Law and Human Behavior*, *32*, 266–278.

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest*, *11*, 89–121.

Walker, J. (2014, December 17). *Are mass shootings becoming more common in the united states?* Retrieved from `http://reason.com`

Walker, L., & Monahan, J. (1988). Social facts: Scientific methodology as legal precedent. *California Law Review*, 877–896.

Walters, G. D., White, T. W., & Denney, D. (1991). The Lifestyle Criminality Screening Form preliminary data. *Criminal Justice and Behavior*, *18*, 406–418.

Warner, S. B. (1923). Factors determining parole from the Massachusetts Reformatory: The report of the director of committee on criminal records and statistics. *Journal of the American Institute of Criminal Law and Criminology*, *14*, 172–207.

Washington State Institute for Public Policy. (1992). *Review of sexual predator program: Community protection research project.*

Webster, C., Harris, G., Rice, M., Cormier, C., & Quinsey, V. (1994). *Violence prediction scheme: Assessing dangerousness in high risk men.* Toronto, Canada: University of Toronto Centre of Criminology.

Webster, C. D., Douglas, K. S., Eaves, D., & Hart, S. D. (1997). HCR-20: Assessing risk of violence, version 2. *Mental Health Law and Policy Institute, Simon Fraser University*.

Webster, C. D., Eaves, D., Douglas, K. S., & Wintrup, A. (1995). *The HCR-20 scheme: The assessment of dangerousness and risk.* Burnaby, Canada: Simon Fraser University and Forensic Psychiatric Services Commission of British Columbia.

Webster, C. D., Martin, M.-L., Brink, J., Nicholls, T. L., & Middleton, C. (2004). *Manual for the Short-Term Assessment of Risk and Treatability (START), Version 1.0, Consultation edition).*

Webster, C. D., Nicholls, T. L., Martin, M.-L., Desmarais, S. L., & Brink, J. (2006). Short-Term Assessment of Risk and Treatability (START): The case for a new structured professional judgment scheme. *Behavioral Sciences and the Law*, *24*, 747–766.

Weinberg, S. K. (1954). Theories of criminality and problems of prediction. *The Journal of Criminal Law, Criminology, and Police Science*, *45*, 412–424.

Weinberger, S. (2012, May 27). *Terrorist 'pre-crime' detector field tested in United States.* Retrieved from `http://www.nature.com`

Weinstein, J. B. (1988). Litigation and statistics. *Statistical Science*, 286–297.

Welch, H. G., Schwartz, L., & Woloshin, S. (2011). *Overdiagnosed: Making people sick in the pursuit of health.* Boston, MA: Beacon Press.

Wenk, E. A., Robison, J. O., & Smith, G. W. (1972). Can violence be predicted? *Crime and Delinquency*, *18*, 393–402.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*, 209–212.

Wis. Stats. ch. 980. (2013).

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science*, *7*, 275–278.

Wolfgang, M. E. (1969). Corrections and the violent offender. *The Annals of the American Academy of Political and Social Science*, *381*, 119–124.

Wollert, R. (2006). Low base rates limit expert certainty when current actuarials are used to identify sexually violent predators: An application of Bayes's theorem. *Psychology, Public Policy, and Law*, *12*, 56–85.

Wong, S. C. P., & Gordon, A. (1999). *Violence Risk Scale*. Saskatoon, Canada: University of Saskatchewan, Department of Psychology.

Yang, M., Liu, Y., & Coid, J. (2010). *Applying neural networks and other statistical models to the classification of serious offenders and the prediction of recidivism.* Ministry of Justice Research Series 6/10.

Yang, M., Wong, S. C. P., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin*, *136*, 740–767.

Yang, S., Mulvey, E. P., Loughran, T. A., & Hanusa, B. H. (2012). Psychiatric symptoms and alcohol use in community violence by persons with a psychotic disorder or depression. *Psychiatric Services*, *63*, 262–269.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*, 32–35.

Zimmerman, M., Galione, J. N., Ruggero, C. J., Chelminski, I., McGlinchey, J. B., Dalrymple, K., & Young, D. (2009). Performance of the mood disorders questionnaire in a psychiatric outpatient setting. *Bipolar Disorders*, *11*, 759–765.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301–320.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*, 561–577.